

LA-UR-05-5219

*Approved for public release;
distribution is unlimited.*

Title: Using Standards in Digital Library Design & Development
(Tutorial)

Author(s):
Jeroen Bekaert
Xiaoming Liu
Herbert Van de Sompel

Submitted to: Joint Conference on Digital Libraries, 2005



Los Alamos National Laboratory, an affirmative action/equal opportunity employer, is operated by the University of California for the U.S. Department of Energy under contract W-7405-ENG-36. By acceptance of this article, the publisher recognizes that the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes. Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy. Los Alamos National Laboratory strongly supports academic freedom and a researcher's right to publish; as an institution, however, the Laboratory does not endorse the viewpoint of a publication or guarantee its technical correctness.

Using Standards in Digital Library Design and Development

Jeroen Bekaert, Xiaoming Liu, and Herbert Van de Sompel
Digital Library Research & Prototyping Team
Research Library, Los Alamos National Laboratory



Using Standards in Digital Library Design and Development
Jeroen Bekaert, Xiaoming Liu & Herbert Van de Sompel
JCDL 2005, June 7, 2005, Denver, CO



overview

- Using Standards in digital library design and development
 - Using MPEG-21 DID to represent Digital Objects
 - Using MPEG-21 DII to identify Digital Objects
 - Using XMLtapes and Internet Archive ARC files to store Digital Objects and constituent datastreams
 - Using OAI-PMH to harvest resources
 - Using the OpenURL Framework to convey Context-Sensitive dissemination requests
 - Using info URI to facilitate the referencing of information assets under the URI allocation
- A use case: the aDORe Digital Object repository at the Research Library of the Los Alamos National Laboratory



Using MPEG-21 DID to represent Digital Objects

Jeroen Bekaert, Xiaoming Liu, and Herbert Van de Sompel
Digital Library Research & Prototyping Team
Research Library, Los Alamos National Laboratory



Using Standards in Digital Library Design and Development
Jeroen Bekaert, Xiaoming Liu & Herbert Van de Sompel
JCDL 2005, June 7, 2005, Denver, CO



what is MPEG?

- Moving Picture Experts Group
 - working group of ISO/IEC in charge of the development of international standards for coded representation of digital audio and video
 - > 300 companies and research facilities: LANL, EPFL, Univ. of Tokyo, Microsoft Research, IBM, Sun Microsystems, Intel, Sony, Motorola, Kodak, Nokia, France Telecom, Hitachi, Mitsubishi, ...
- From MPEG-1 to MPEG-21
 - MPEG-1 (ISO/IEC 11172) CD-ROM, MP-3 (1992)
 - MPEG-2 (ISO/IEC 13818) Digital TV, DVD (1994)
 - MPEG-4 (ISO/IEC 14496) Object-based video compression
 - MPEG-7 (ISO/IEC 15938) Metadata for multimedia content
 - MPEG-21 (ISO/IEC 21000) Multimedia Framework

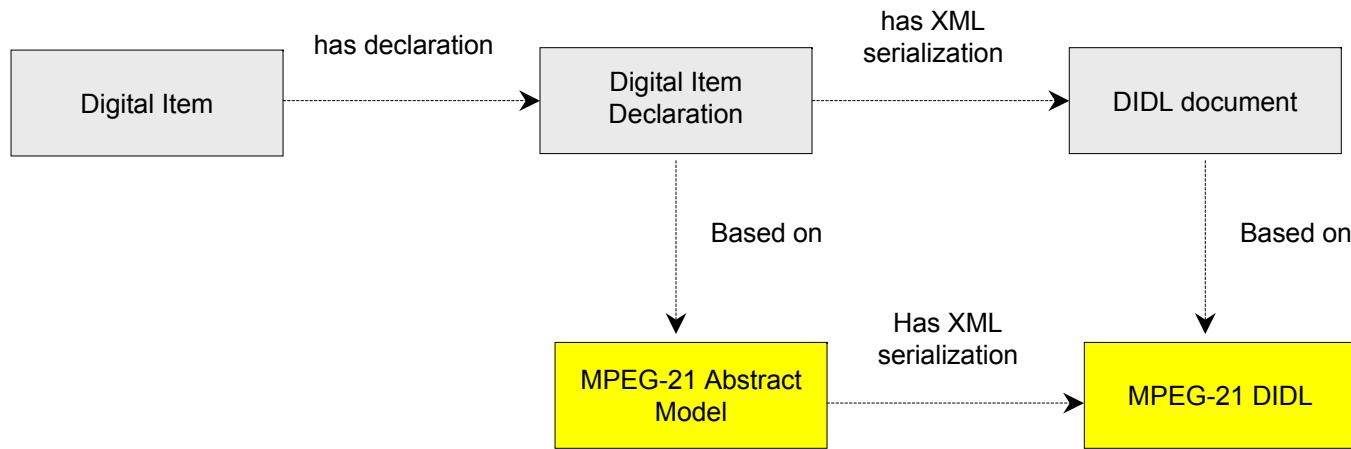
ISO/IEC 21000: MPEG-21

- ‘MPEG-21’s vision: *‘to define a set of tools to support the usage of content. In order to facilitate interoperability within a domain or between domains, tools may be used independently or selectively in combination’*
 - Applicable to Digital Libraries domain
 - Ability to accomodate any media type and genre
- MPEG-21 is modular (17 parts):
 - Part 2: MPEG-21 DID – XML representation of digital objects
 - Part 3: MPEG-21 DII – identification of digital objects
 - Part 4: MPEG-21 IPMP – enforcement of rights expressions
 - Part 5: MPEG-21 REL – declaration of rights expressions
 - Part 6: MPEG-21 RDD – dictionary of rights related terms (<indecs)
 - Part 7: MPEG-21 DIA – (transcoding based on) contextual information
 - Part 16: MPEG-21 BF – binary representation of digital objects

ISO/IEC 21000-2: MPEG-21 DID

- ISO/IEC 21000-2: MPEG-21 Digital Item Declaration
 - first edition: ISO/IEC – March 2004
 - second edition: ISO/IEC – April 2005 (registered for formal approval – freely available)
- Scope: representation and packaging of Digital Items. A Digital Item is:
 - a compound object containing one or more constituting datastreams and key-metadata
 - aka Content Information object (ISO OAIS RM)
 - aka Digital Object (Kahn/Wilensky)
 - aka Asset (LANL aDORe environment)
- 2 distinct sections:
 - MPEG-21 DID Abstract Model
 - abstract data model for declaring Digital Items
 - result of mapping a Digital Item to the model = Digital Item Declaration
 - MPEG-21 Digital Item Declaration Language (MPEG-21 DIDL)
 - serialization of the data model in XML
 - result of serializing/packaging a Digital Item Declaration = DIDL document

Abstract Model & MPEG-21 DIDL



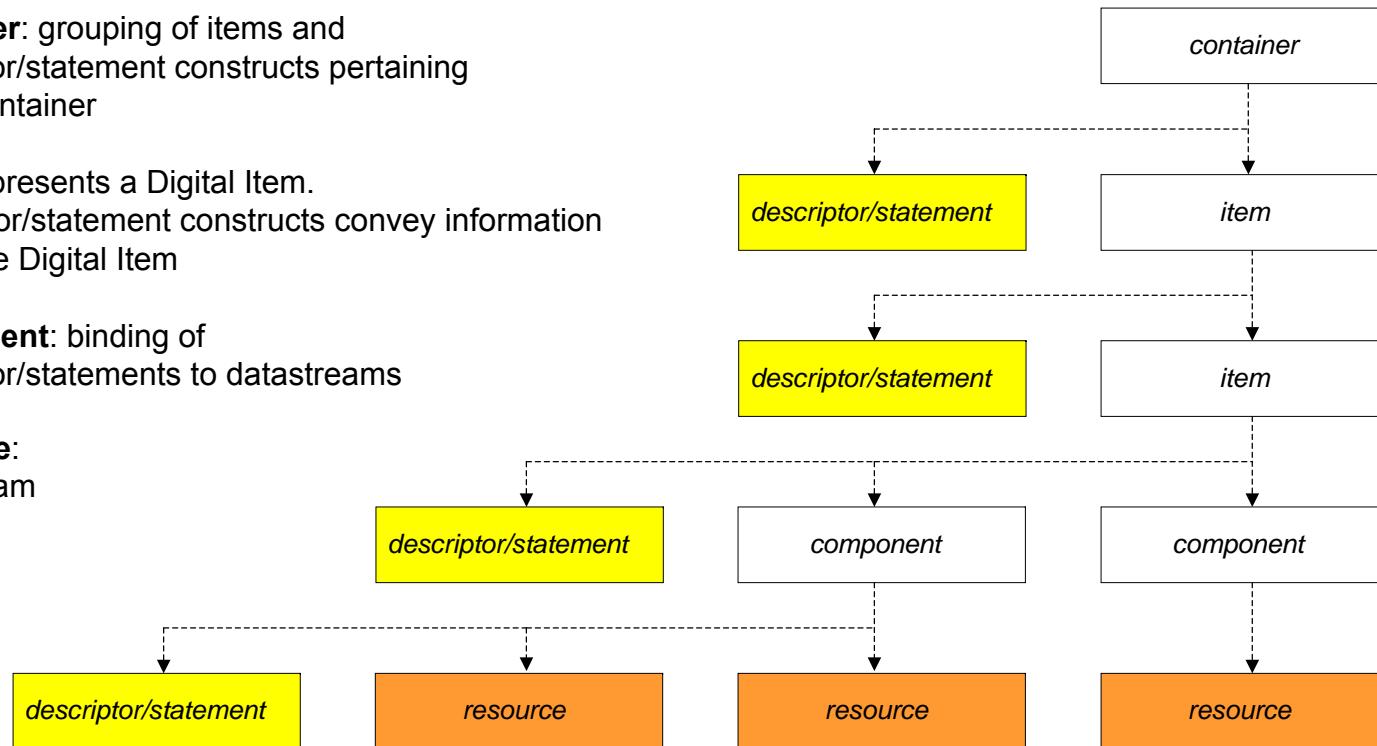
Abstract Model: basic entities

container: grouping of items and descriptor/statement constructs pertaining to the container

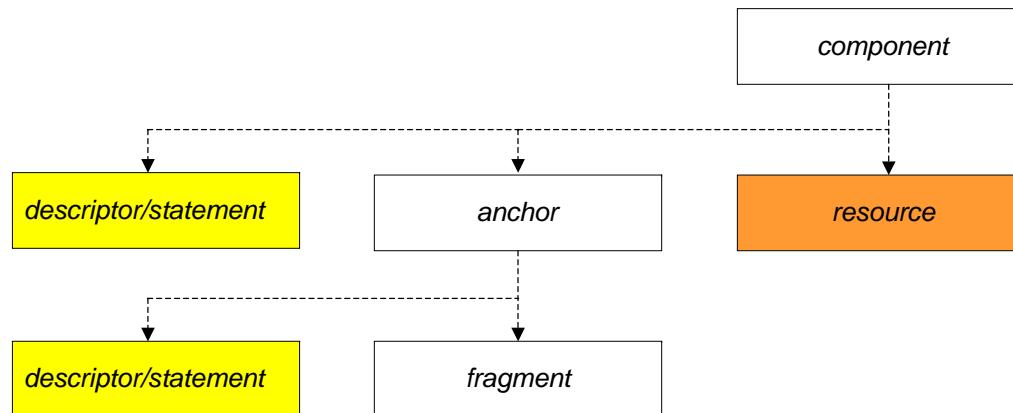
item: represents a Digital Item.
Descriptor/statement constructs convey information about the Digital Item

component: binding of descriptor/statements to datastreams

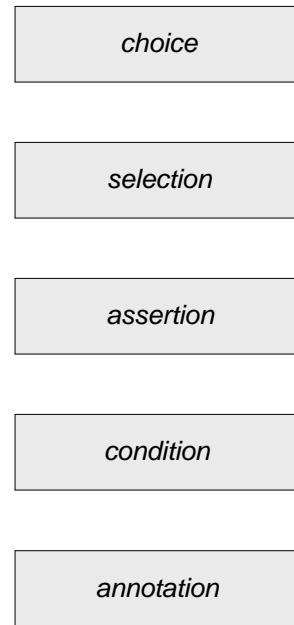
resource:
datastream



Abstract Model: *anchor*



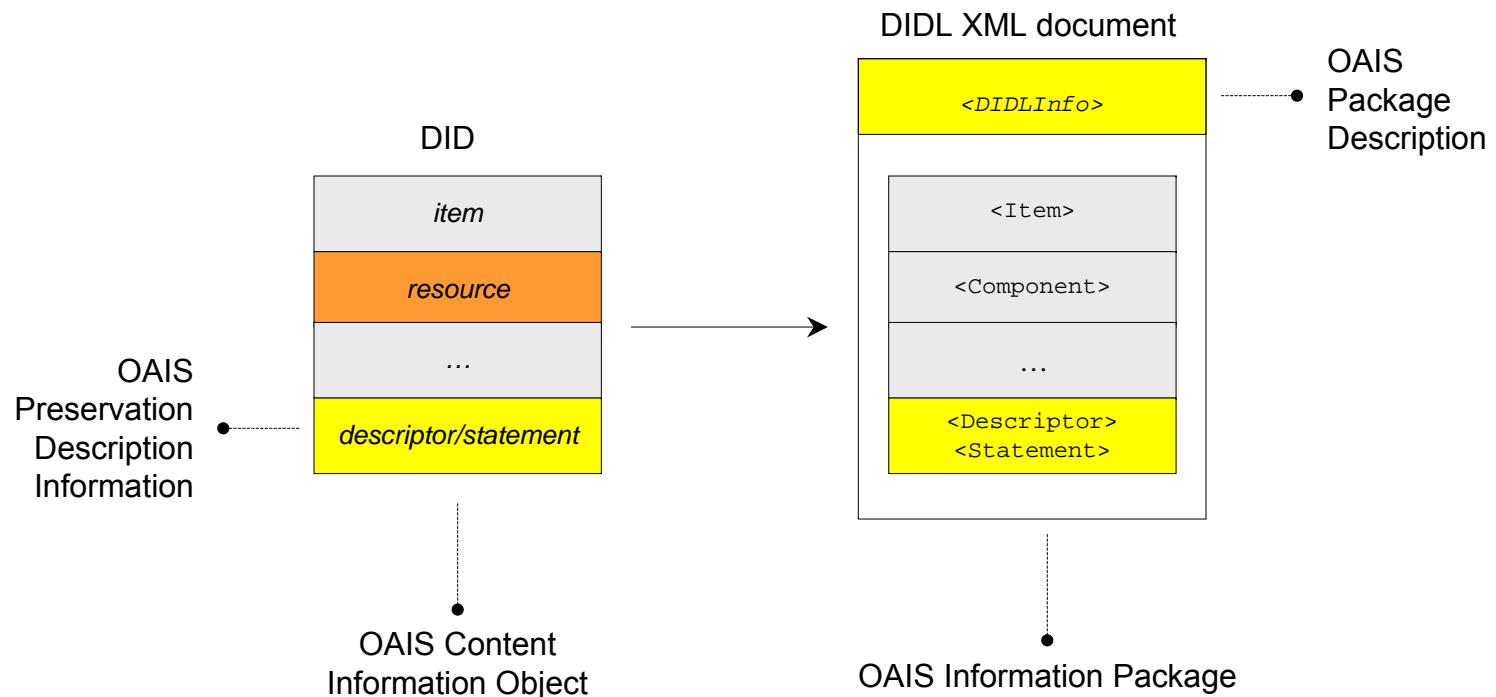
Abstract Model: *choice, condition, annotation*



Using Standards in Digital Library Design and Development
Jeroen Bekaert, Xiaoming Liu & Herbert Van de Sompel
JCDL 2005, June 7, 2005, Denver, CO



MPEG-21 DIDL



MPEG-21 DIDL: data provision techniques

		<DIDLInfo>	<Descriptor> <Statement>	<Resource>
By value	character data		■	■
	well-formed XML	□	■	□
	base64 encoded data		■	■
By Reference	unencoded data		□	■

DIDL XML serialization

secondary information

- Secondary information pertaining to DIDL documents
 - via `DIDLInfo` element
 - via attributes attached to `DIDL` element
- Secondary information pertaining to Digital Items
 - via Descriptor/Statement constructs
 - via attributes attached to Container/Item/Component
 - MPEG-21 pre-defined uses
 - Identification information – MPEG-21 DII (ISO/IEC 21000-3)
 - Rights information – MPEG-21 REL (ISO/IEC 21000-5)
 - community/application specific uses
 - aDORe specific Descriptor/Statement constructs
 - OAI-PMH specific Descriptor/Statement constructs

secondary information

MPEG-21 REL: rights information (content oriented)

```
<didl:Item>
...
<didl:Descriptor>
    <didl:Statement mimeType="text/xml; charset=UTF-8">
        <r:license xmlns:r="urn:mpeg:mpeg21:2003:01-REL-R-NS">
            <!-- optionally, specific rights can be added here.-->
            <r:otherInfo>
                <dc:rights xmlns:dc="http://purl.org/dc/elements/1.1/">
                    Copyright2003; American Physical Society</dc:rights>
            </r:otherInfo>
        </r:license>
    </didl:Statement>
</didl:Descriptor>
...
</didl:Item>
```

MPEG-21 r:license



Using Standards in Digital Library Design and Development
Jeroen Bekaert, Xiaoming Liu & Herbert Van de Sompel
JCDL 2005, June 7, 2005, Denver, CO



secondary information

aDORe: creation datetime information (package & content oriented)

```
<didl:DIDL diext:"DIDcreated=2004-12-04T01:01:01Z"  
    xmlns:didl="urn:mpeg:mpeg21:2002:02-DIDL-NS"  
    xmlns:diext= "http://library.lanl.gov/2004-04/STB-RL/DIEXT">  
    <didl:Item>  
        ...  
    </didl:Item>  
</didl:DIDL>
```

aDORe diext:DIDcreated

```
<didl:Component id="#uuid-a0577072-992a-11d8-b3f1-fc62348d6ec0">  
    ...  
    <didl:Descriptor>  
        <didl:Statement mimeType="text/xml; charset=UTF-8">  
            <diadm:Admin xmlns:diadm="http://library.lanl.gov/2004-01/STB-RL/DIADM">  
                <dcterms:created xmlns:dcterms="http://purl.org/dc/terms/">  
                    2004-05-18T15:43:33Z</dcterms:created>  
                </diadm:Admin>  
            </didl:Statement>  
        </didl:Descriptor>  
        ...  
</didl:Component>
```

aDORe diadm:Admin



secondary information

aDORe: W3C XML signature constructs (datastream oriented)

```
<didl:Component id="#uuid-a0577072-992a-11d8-b3f1-fc62348d6ec0">
...
<didl:Descriptor>
  <didl:Statement mimeType="text/xml; charset=UTF-8">
    <dsig:Signature xmlns:dsig="http://www.w3.org/2000/09/xmldsig#">
      <dsig:SignedInfo>
        <dsig:CanonicalizationMethod
          Algorithm="http://www.w3.org/TR/2001/REC-xml-c14n-20010315"/>
        <dsig:SignatureMethod Algorithm="http://www.w3.org/2000/09/xmldsig#rsa-sha1"/>
        <dsig:Reference URI="#uuid-a0577072-992a-11d8-b3f1-fc62348d6ec0">
          ...
        </dsig:Reference>
      </dsig:SignedInfo>
      <dsig:SignatureValue>dRaxVQYPMd0vfzkbstaG8taNTtJA9sF9ze3/xW6AeW9KCguijhpmG2kAuDJhe
+EA7X0uNf59UIanLlMiGh3+ROzctwy00z8vbKqjGxYU=</dsig:SignatureValue>
    </dsig:Signature>
  </didl:Statement>
</didl:Descriptor>
...
</didl:Component>
```

W3C dsig:Signature

secondary information

aDORe: W3C XML signature constructs (package oriented)

```
<didl:DIDL xmlns:didl="urn:mpeg:mpeg21:2002:02-DIDL-NS">
  <didl:DIDLInfo>
    <dsig:Signature xmlns:dsig="http://www.w3.org/2000/09/xmldsig#">
      <dsig:SignedInfo>
        <dsig:CanonicalizationMethod
          Algorithm="http://www.w3.org/TR/2001/REC-xml-c14n-20010315"/>
        <dsig:SignatureMethod Algorithm="http://www.w3.org/2000/09/xmldsig#rsa-sha1"/>
        <dsig:Reference>
          ...
        </dsig:Reference>
      </dsig:SignedInfo>
      <dsig:SignatureValue>dRaxVQYPMd0vfzbstaG8taNTtJA9sF9ze3/xW6AeW9KCguIjHpmG2kAuDJhe
+EA7X0uNf59UIanLlMiGh3+ROzctwyO0z8vbKqjGxYU=</dsig:SignatureValue>
    </dsig:Signature>
  </didl:DIDLInfo>
  <didl:Item>
    ...
  </didl:Item>
</didl:DIDL>
```

W3C dsig:Signature



Using Standards in Digital Library Design and Development
Jeroen Bekaert, Xiaoming Liu & Herbert Van de Sompel
JCDL 2005, June 7, 2005, Denver, CO



secondary information

MPEG-21 DII: identification information (content oriented)

```
<didl:Item>
  <didl:Descriptor>
    <didl:Statement mimeType="text/xml; charset=UTF-8">
      <dii:Identifier xmlns:dii="urn:mpeg:mpeg21:2002:01-DII-NS">
        urn:isbn:0-395-36341-1</dii:Identifier>
      </didl:Statement>
    </didl:Descriptor>
  ...
</didl:Item>
```

MPEG-21 dii:Identifier



Using Standards in Digital Library Design and Development
Jeroen Bekaert, Xiaoming Liu & Herbert Van de Sompel
JCDL 2005, June 7, 2005, Denver, CO



secondary information

MPEG-21 DII: identification information (content oriented)

```
<didl:Item>
  <didl:Descriptor>
    <didl:Statement mimeType="text/xml; charset=UTF-8">
      <dii:Identifier xmlns:dii="urn:mpeg:mpeg21:2002:01-DII-NS">
        urn:mpegRA:mpeg21:dii:doi:10.1000/123456789</dii:Identifier>
      </didl:Statement>
    </didl:Descriptor>
    <didl:Descriptor>
      <didl:Statement mimeType="text/xml; charset=UTF-8">
        <dii:relatedIdentifier xmlns:dii="urn:mpeg:mpeg21:2002:01-DII-NS"
          relationshipType="urn:mpeg:mpeg21:2002:01-RDD-NS:IsAbstractionOf">
          urn:mpegRA:mpeg21:dii:iswc:T-034.524.680-1</dii:relatedIdentifier>
        </didl:Statement>
      </didl:Descriptor>
      ...
    </didl:Item>
```

MPEG-21 dii:RelatedIdentifier

secondary information

MPEG-21 DID: identification information (package oriented)

```
<didl:DIDL DIDLDocumentId="info:lanl-repo/i/00002cb8-c477"  
    xmlns:didl="urn:mpeg:mpeg21:2002:02-DIDL-NS">  
    <didl:Item>  
        ...  
    </didl:Item>  
</didl:DIDL>
```

MPEG-21 didl:DIDLDocumentId



Using Standards in Digital Library Design and Development
Jeroen Bekaert, Xiaoming Liu & Herbert Van de Sompel
JCDL 2005, June 7, 2005, Denver, CO



sample Digital Object

	Type	MIME	identifier
Digital Object	scholarly paper	N/A	DOI
Constituent Datastream 1	metadata record	application/xml	PMID
Constituent Datastream 2	fulltext file	application/pdf	–

OAIS PACKAGE PERSPECTIVE

OAIS CONTENT PERSPECTIVE





software & reading

- Software:
 - MPEG-21 DID: Reference Software (ISO/IEC 21000-8 - registered for formal approval)
 - MPEG-21 DID: Perl-based MPEG-21 DID writer, to be released in CPAN (developed by LANL)
 - MPEG-21 DID: Java-based MPEG-21 DID writer, based on Apache XMLBeans, ongoing work (developed by LANL and Ghent University)
 - XML Signatures: Apache XML Security,
<http://xml.apache.org/security/Java/index.html>.
- Readings (see also supplied CD-rom):
 - ISO/IEC 21000-2. MPEG-21 Digital Item Declaration specification.
<http://www.iso.org>
 - ISO/IEC 21000-3. MPEG-21 Digital Item Identification specification.
<http://www.iso.org>

File-based storage of Digital Objects and constituent datastreams: XMLtapes and Internet Archive ARC files

Jeroen Bekaert, Xiaoming Liu, and Herbert Van de Sompel
Digital Library Research & Prototyping Team
Research Library, Los Alamos National Laboratory



Using Standards in Digital Library Design and Development
Jeroen Bekaert, Xiaoming Liu & Herbert Van de Sompel
JCDL 2005, June 7, 2005, Denver, CO



introduction

- Storage of XML-based OAIS AIPs (consisting of a variety of constituent datastreams and secondary information pertaining to a Digital Object)
- Existing approaches:
 - storage of the XML-wrapper documents as individual files in a file system:
 - Poor access performance
 - Poor backup performance
 - storage of the XML-wrapper documents in SQL or native XML databases
 - Long term? Data are dependent on the underlying system
 - storage of the XML-wrapper documents by concatenating many such documents into a single file such as tar or zip
 - Not XML aware, hence, no use of off-the-shelf XML tools
 - Increasing storage space (base64-encoding of the constituent datastreams)

XMLtape/ARC solution

- Storage of compound objects (independent of the choice of complex object format)
- Write once - Read many: Files remain stable while indexs mechanisms can change when technologies evolve
- Two interconnected file-based storage mechanisms:
 - **XMLtapes**: File storage of XML-based representations of Digital Objects
 - **ARC files**: File storage of constituent datastreams of Digital Objects
- The ARC files are interconnected with one or more XMLtapes during the ingestion process
- A protocol-based access mechanism is introduced
 - XMLTape is exposed as an autonomous OAI-PMH repository
 - ARC file is exposed as an OpenURL Resolver

XMLtape

- An XML file that concatenates the XML-based representations of multiple Digital Objects
- Structure is defined by an XML Schema
 - tape-level administrative section
 - e.g. containing processing related information)
 - concatenation of records, each of which consists of:
 - record-level administrative section
 - identifier and datestamp of the contained record
 - other record-level administrative information
 - a record (can be from any XML Namespace)
- The XMLtape is a valid and well-formed XML file
- Developed at LANL



XMLtape

```
<?xml version="1.0" encoding="UTF-8"?>
<ta:tape xmlns:ta="http://library.lanl.gov/2005-01/STB-RL/tape/"
          xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
    <ta:tape-admin>
        ...
    </ta:tape-admin>
    <ta:tape-record>
        <ta:tape-record-admin>
            <ta:identifier>oai:aps.org:PhysRevA.71.040101</ta:identifier>
            <ta:date>2005-03-29T04:31:22Z</ta:date>
            <ta:record-admin>
                ...
            </ta:record-admin>
        </ta:tape-record-admin>
        <ta:record>
            <didl:DIDL>...</didl:DIDL>
        </ta:record>
    </ta:tape-record>
</ta:tape>
```

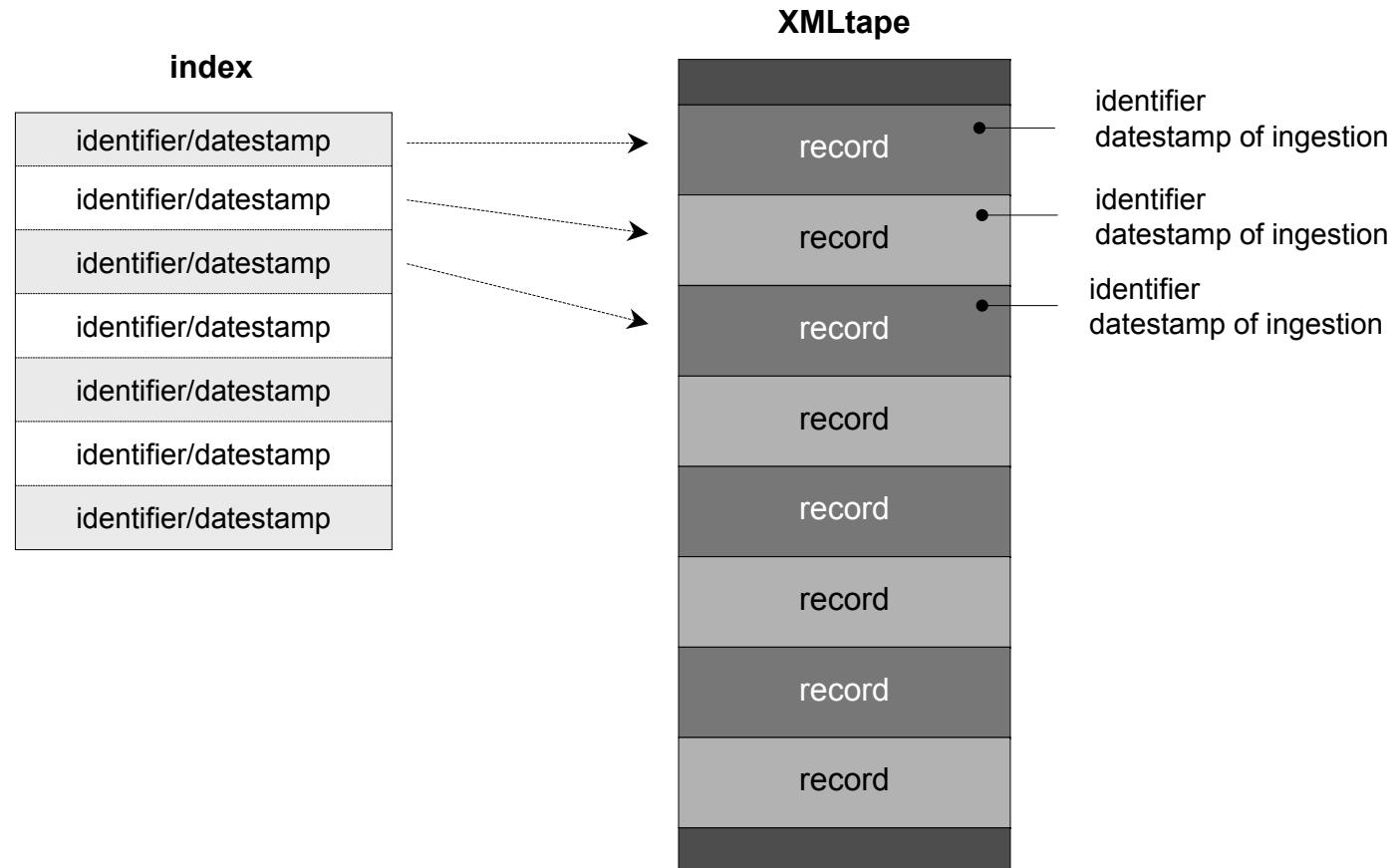
aDORe ta:tape



Using Standards in Digital Library Design and Development
Jeroen Bekaert, Xiaoming Liu & Herbert Van de Sompel
JCDL 2005, June 7, 2005, Denver, CO



XMLtape



Internet Archive ARC file

- Concatenation of binary files
- Designed and used by the Internet Archive (Wayback machine)
 - > 400 TB web data
- Under revision by the International Internet Preservation Consortium (IIPC)
- The ARC file format is structured as follows:
 - file header that provides administrative information about the ARC file itself
 - a sequence of document records, consisting of:
 - a header line containing some, mainly crawl-related, metadata.
 - URI of the crawled document
 - timestamp of acquisition of the data
 - size of the data block
 - a response to a protocol request such as an HTTP GET

Internet Archive ARC file

```
filedesc://IA-001102.arc 0.0.0.0 19960923142103 text/plain 200 - - 0
IA-001102.arc 122
2 0 Alexa Internet
URL IP-address Archive-date Content-type Result-code Checksum
Location Offset Filename Archive-length

http://www.dryswamp.edu:80/index.html 127.10.100.2 19961104142103
text/html 200 fac069150613fe55599cc7fa88aa089d - 209 IA-001102.arc 202
HTTP/1.0 200 Document follows
Date: Mon, 04 Nov 1996 14:21:06 GMT
Server: NCSA/1.4.1
Content-type: text/html Last-modified: Sat,10 Aug 1996 22:33:11 GMT
Content-length: 30
<HTML>
Hello World!!!
</HTML>
```

sample ARC file

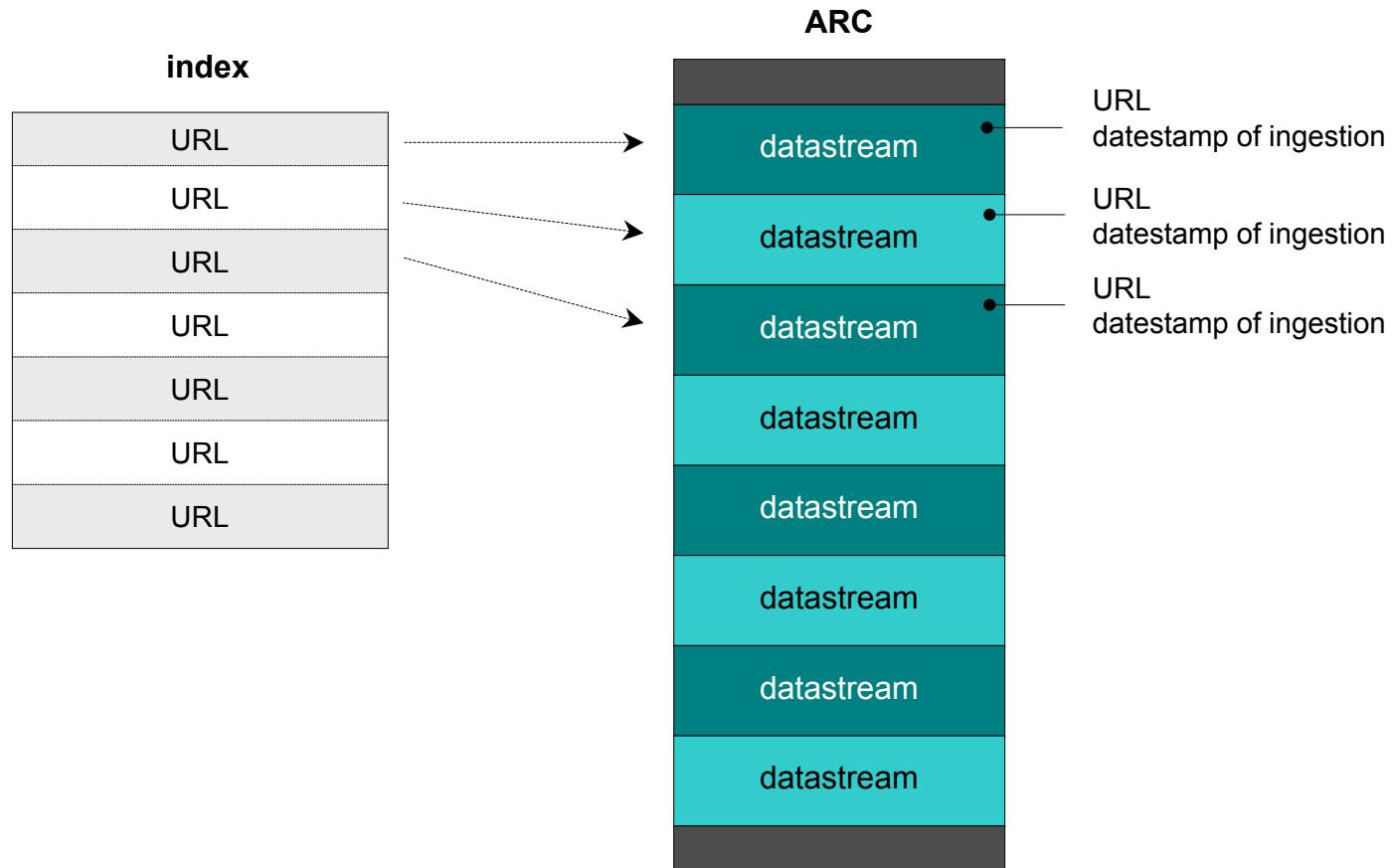


RESEARCH
LIBRARY

Using Standards in Digital Library Design and Development
Jeroen Bekaert, Xiaoming Liu & Herbert Van de Sompel
JCDL 2005, June 7, 2005, Denver, CO



Internet Archive ARC file



Internet Archive ARC file – ongoing revision

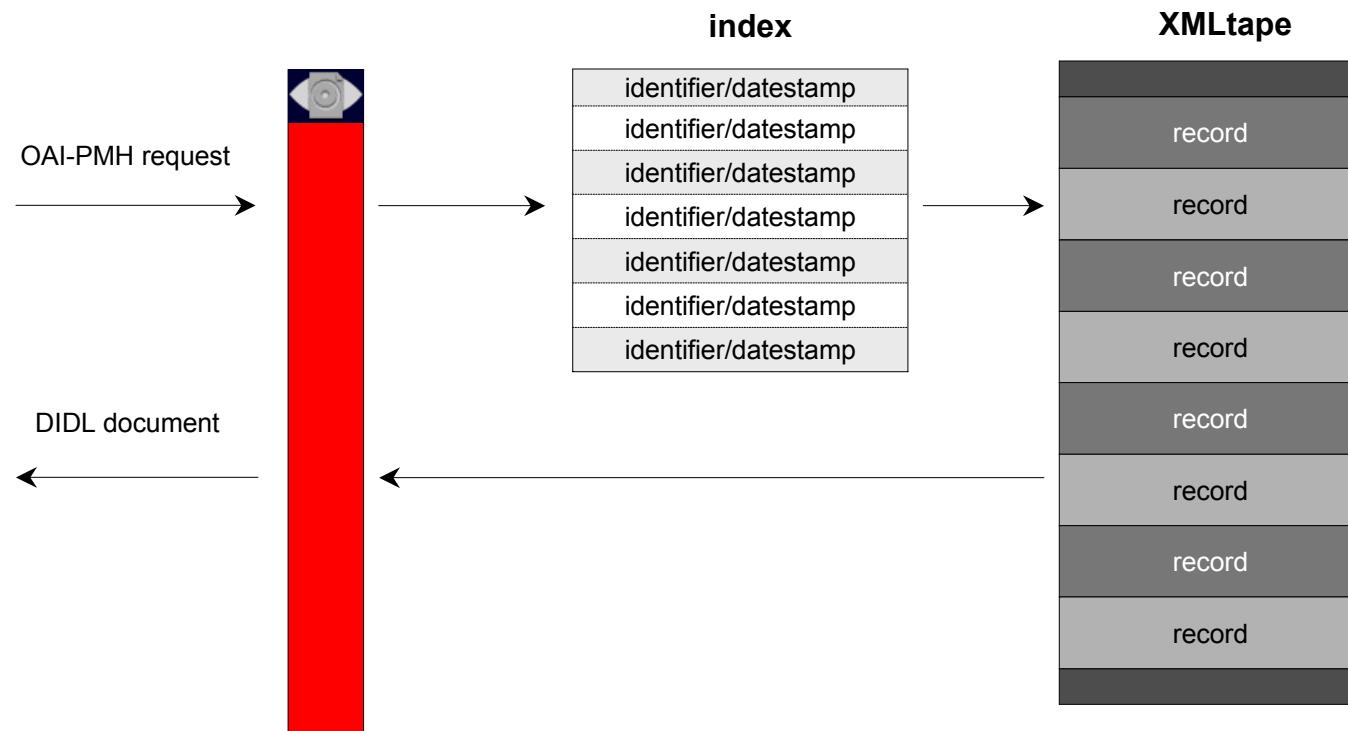
- In collaboration with IIPC, Internet Archive, and the national libraries of a dozen countries.
- Generalize the older format to support harvesting, display, and exchange needs of the archiving domain.
- Accommodate secondary information, such as assigned metadata, duplicate detection, and transformation.
- Useful to more general applications than web archiving



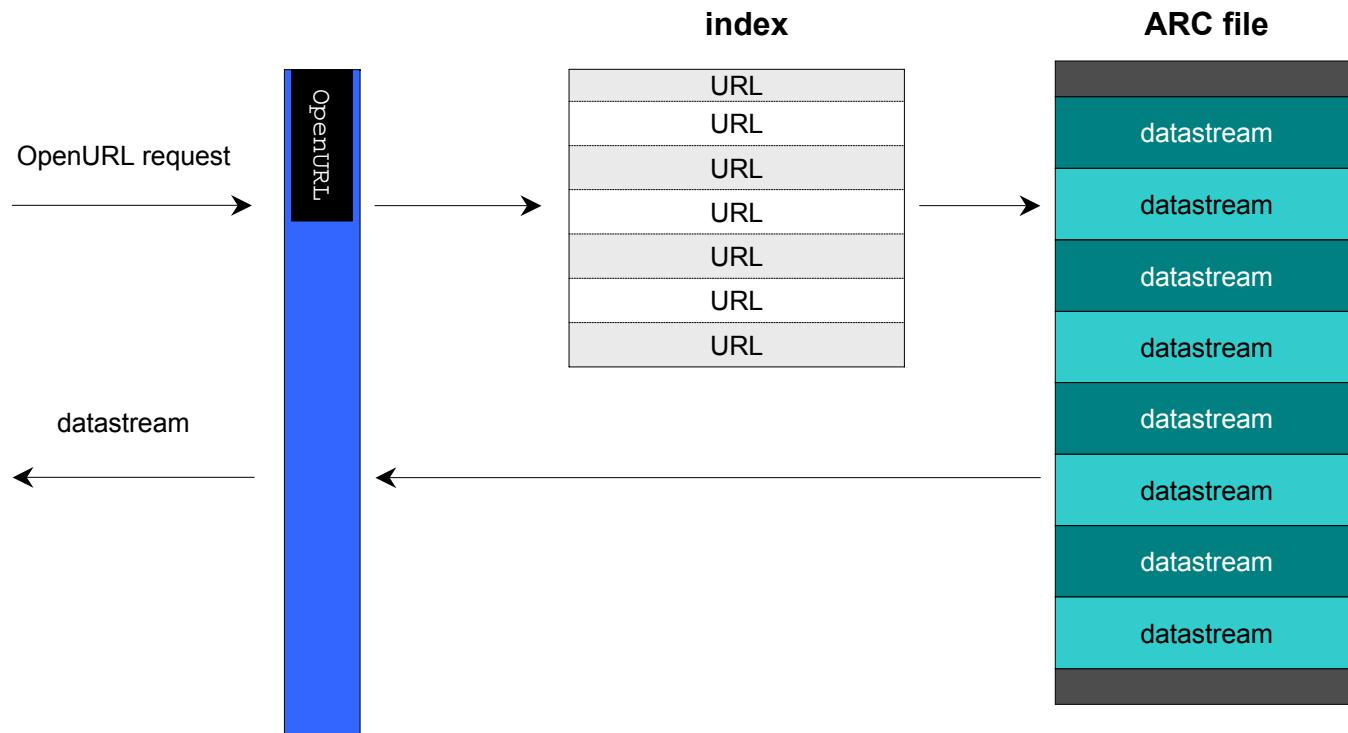
associating XMLtape and ARC Files

- Both formats can be used separately.
 - XMLtapes: storing well-formed XML data.
 - ARC files: storing of binary data.
- Associating XMLtapes and ARC files:
 - A Digital Object is represented and packaged using a Complex Object format (e.g. MPEG-21 DID)
 - The resulting package (DIDL document) is stored in an XMLtape
 - Constituent datastreams of the Digital Object are provided By-Reference (using the `ref` attribute of the `Resource` element in MPEG-21 DID)
 - The constituent datastreams are stored in one or more Internet Archive ARC files
 - URI of an ARC record = Datastream Identifier
 - The value of the network location of the constituent datastream is compliant with the NISO OpenURL Framework
 - `baseURL(ARCfile Identifier)?`
 - `url_ver=Z39.88-2004 &`
 - `rft_id=Datastream Identifier`
- In addition, ARC files are connected with an XMLtape by including the ARCfile Identifiers in the XMLtape-level administrative section.

XMLtape as OAI-PMH repository



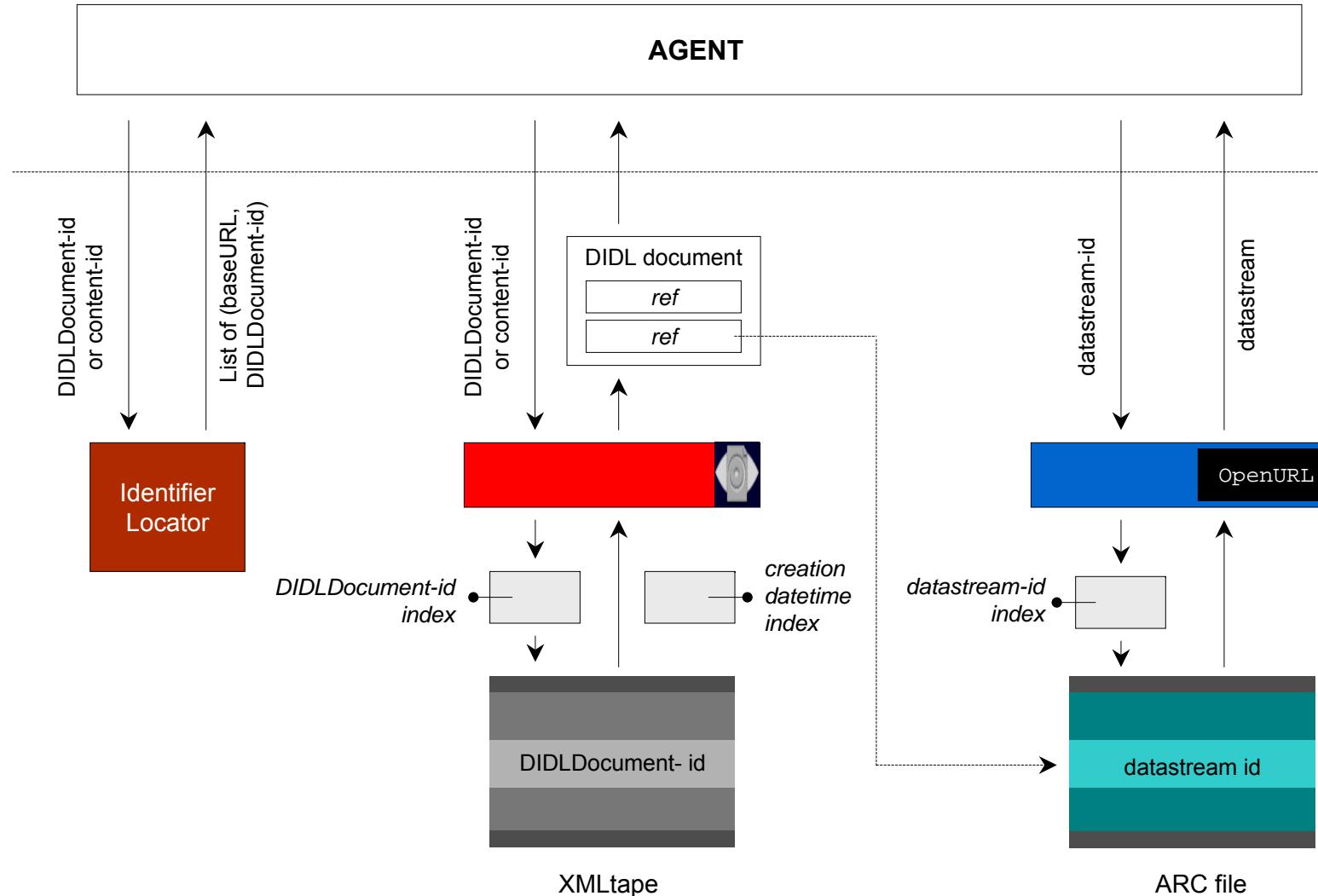
ARC file as OpenURL Resolver



Using Standards in Digital Library Design and Development

Jeroen Bekaert, Xiaoming Liu & Herbert Van de Sompel
JCDL 2005, June 7, 2005, Denver, CO





Conclusion

- The file-based approach is inherently simple, and reduces dependency on database system.
- The disconnection of the indexes allows retaining the files over time, while the indexes can be created using other techniques as technologies evolve.
- The protocol-based nature of the access increases the flexibility in light of evolving technologies as it introduces another layer of abstraction.
- The XMLtape approach is inspired by the ARC file format, but provides several additional attractive features. Off-the-shelf XML tools can be used to parse/validate an XMLtape.



Software & readings

- Software - ARC files:
 - Heritrix: the internet archive's open-source, extensible, web-scale, archival-quality web crawler project. <http://crawler.archive.org/>
 - NetArchive.dk: a project that plans for the preservation of Denmark's cultural heritage on the internet for future generations. <http://www.netarchive.dk/>
 - Many other tools: <http://archive-access.sourceforge.Net>
- Software - XMLtapes:
 - Perl tool, YAR (LANL), <http://yar.sourceforge.net>
Java tool (LANL), to be released, based on OCLC oaicat, and Berkeley DB java Edition.
- Readings:
 - File-based storage of Digital Objects and constituent datastreams: XMLtapes and Internet Archive ARC files. <http://arxiv.org/pdf/cs.DL/0503016>
 - Arc File Format. <http://www.archive.org/web/researcher/ArcFileFormat.php>

OAI-PMH for Resource Harvesting

Jeroen Bekaert, Xiaoming Liu, and Herbert Van de Sompel
Digital Library Research & Prototyping Team
Research Library, Los Alamos National Laboratory



Using Standards in Digital Library Design and Development
Jeroen Bekaert, Xiaoming Liu & Herbert Van de Sompel
JCDL 2005, June 7, 2005, Denver, CO



Resource Harvesting: use cases

- Discovery: use content itself in the creation of services
 - search engines that make full-text searchable
 - citation indexing systems that extract references from the full-text content
 - browsing interfaces that include thumbnail versions of high-quality images from cultural heritage collections
- Preservation:
 - periodically transfer digital content from a data repository to one or more trusted digital repositories
 - trusted digital repositories need a mechanism to automatically synchronize with the originating data repository

Resource Harvesting: use cases

- Discovery:
 - Institutional Repository & Digital Library Projects: UK JISC, DARE, DINI
 - Web search engines: competition for content (cf Google Scholar)
- Preservation:
 - Institutional Repository & Digital Library Projects: UK JISC, DARE, DINI
 - Library of Congress NDIIP Archive Export/Ingest

OAI-PMH is well-established.
Can OAI-PMH be used for Resource Harvesting?

Existing OAI-PMH based approaches

- Typical scenario:
 - an OAI-PMH harvester harvests Dublin Core records from the OAI-PMH repository.
 - the harvester analyzes each Dublin Core record, extracting dc.identifier information in order to determine the network location of the described resource.
 - a separate process, out-of-band from the OAI-PMH, collects the described resource from its network location.

Existing OAI-PMH based approaches : issue 1

- Locating the resource based on information provided in `dc.identifier`
 - `dc.identifier` is used to convey a variety of identifiers: (simultaneously) URL DOI, bibliographic citation, ...
 - not expressive enough to distinguish between identifier, locator.
 - several dereferencing attempts required
 - URI provided in `dc.identifier` is commonly that of a bibliographic “splash page”
 - How to know it is a bibliographic “splash page”, not the resource?
 - If it is a bibliographic “splash page”, where is the resource?

Existing OAI-PMH based approaches : issue 2

- Using the OAI-PMH datestamp of the Dublin Core record to trigger incremental harvesting:
 - datestamp of DC record does not necessarily change when resource changes

		DC record datestamp no change	DC record datestamp change
		no metadata update	metadata update
no resource update	OK	unnecessary resource download	
resource update	missed resource update	OK	

Existing OAI-PMH based approaches : conventions

- Conventions address Issue 1; Issue 2 can not really be addressed.
- First `dc.identifier` is locator of the resource
 - what if the resource is not digital?
- Use of `dc.format` and/or `dc.relation` to convey locator



Existing OAI-PMH based approaches : conventions

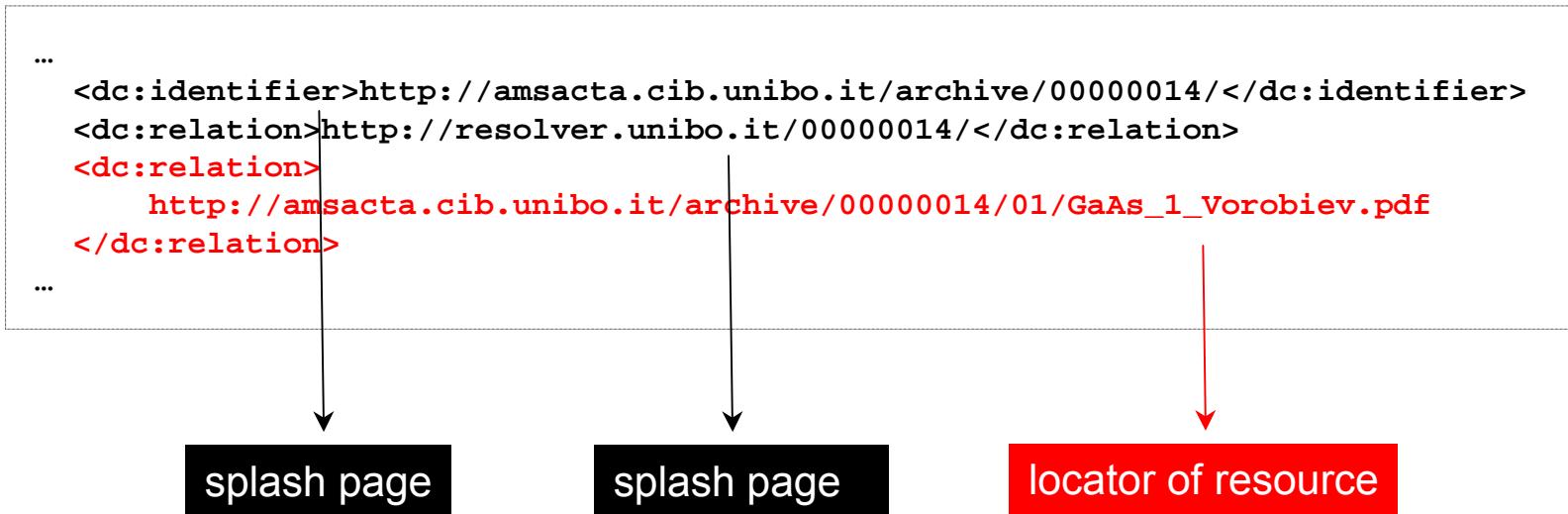
```
<oai_dc:dc>
  <dc:title>A Simple Parallel-Plate Resonator Technique for Microwave.
    Characterization of Thin Resistive Films</dc:title>
  <dc:creator>Vorobiev, A.</dc:creator>
  <dc:subject>ING-INF/01 Elettronica</dc:subject>
  <dc:description>A parallel-plate resonator method is proposed for
    non-destructive characterisation of resistive films used in
    microwave integrated circuits. A slot made in one ... </dc:description>
  <dc:publisher>Microwave engineering Europe</dc:publisher>
  <dc:date>2002</dc:date>
  <dc:type>Documento relativo ad una Conferenza o altro Evento</dc:type>
  <dc:type>PeerReviewed</dc:type>
  <dc:identifier>http://amsacta.cib.unibo.it/archive/00000014/</dc:identifier>
  <dc:format>
    http://amsacta.cib.unibo.it/archive/00000014/01/GaAs\_1\_Vorobiev.pdf
  </dc:format>
</oai_dc:dc>
```



Existing OAI-PMH based approaches : conventions



Existing OAI-PMH based approaches : conventions



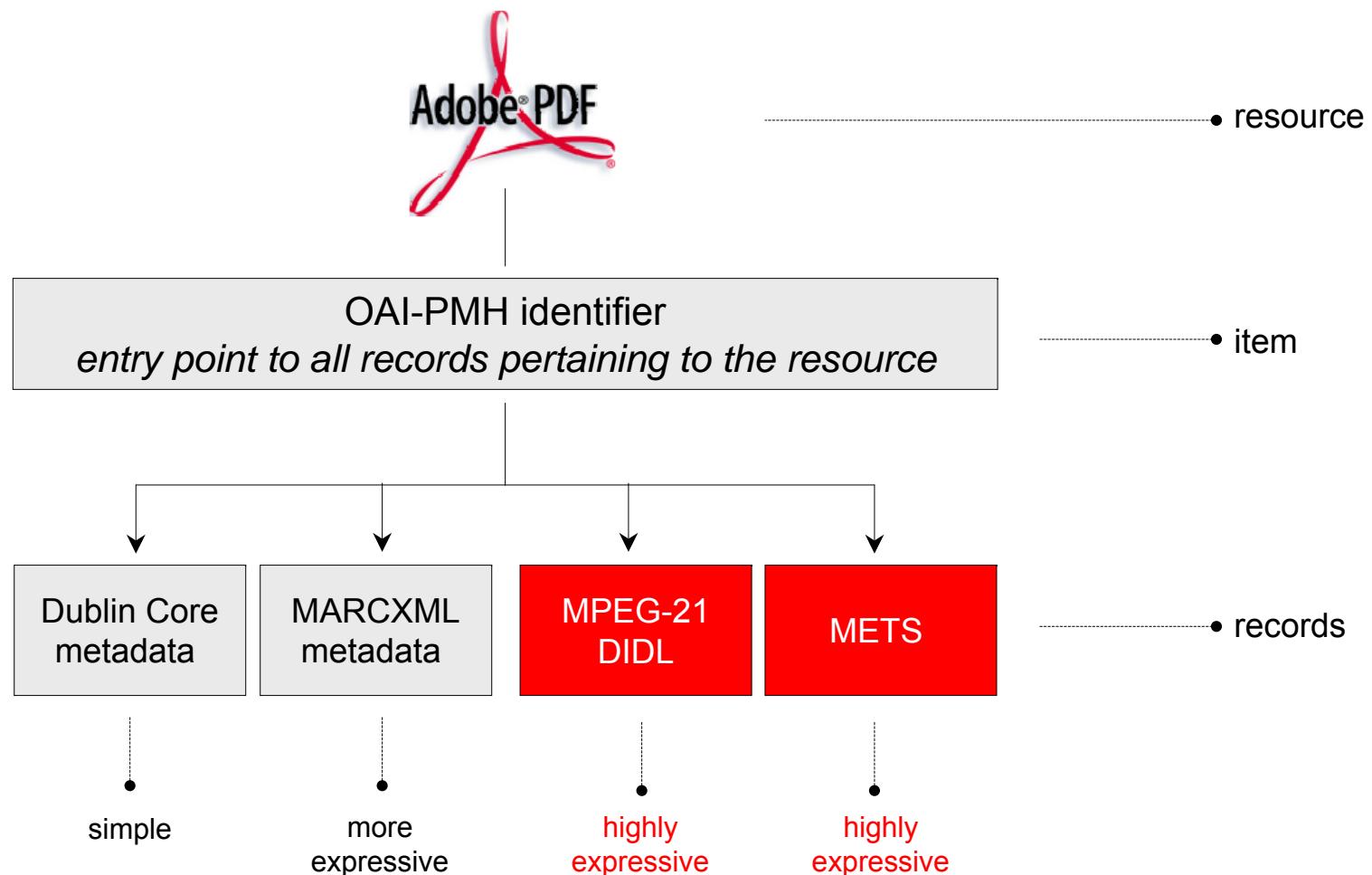
Existing OAI-PMH based approaches : other attempts

- dc.identifier leads to splash page & splash page contains special purpose XHTML link to resource(s)
 - what if there is no splash page?
 - how does a harvester know he is in this situation?
- OA-X: protocol extension
 - OK in local context
 - strategic problem to generalize
 - how to consolidate with OAI-PMH data model
- Qualified Dublin Core
 - could bring expressiveness to distinguish between locator & identifier
 - but what with datestamp issue?

Proposed OAI-PMH based approach

- Use XML-based metadata formats that were specifically created for representation of digital objects:
 - Complex Object Formats as OAI-PMH metadata formats
 - MPEG-21 DIDL, METS, XFDU, IMS-CP Manifest, ...

OAI-PMH data model



Complex Object Formats : characteristics

- Representation of a digital object by means of a wrapper XML document
- Represented resource can be:
 - simple digital object (consisting of a single datastream)
 - compound digital object (consisting of multiple datastreams)
- Unambiguous approach to convey identifiers of the digital object and its constituent datastreams
- Include datastream:
 - By-Value: embedding of base64-encoded datastream
 - By-Reference: embedding network location of the datastream
 - not mutually exclusive; equivalent
- Include a variety of secondary information
 - By-Value
 - By-Reference
 - descriptive metadata, rights information, technical metadata, ...

```

<didl:DIDL>
<didl:Item>
  <didl:Descriptor>
    <didl:Statement mimeType="text/xml; charset=UTF-8">
      <dii:Identifier>
        http://amsacta.cib.unibo.it/archive/00000014/
      </dii:Identifier>
    </didl:Statement>
  </didl:Descriptor>
  <didl:Descriptor>
    <didl:Statement mimeType="text/xml; charset=UTF-8">
      <oai_dc:dc>
        <dc:title>A Simple Parallel-Plate Resonator Technique for
          Microwave. Characterization of Thin Resistive Films
        </dc:title>
        <dc:creator>Vorobiev, A.</dc:creator>
        <dc:identifier>
          http://amsacta.cib.unibo.it/archive/00000014/<dc:identifier>
          <dc:format>application/pdf</dc:format>
        </oai_dc:dc>
      </didl:Statement>
    </didl:Descriptor>
    <didl:Component>
      <didl:Resource mimeType="application/pdf"
        ref="http://amsacta.cib.unibo.it/archive/00000014/01/GaAs_1_Vorobiev.pdf"/>
    </didl:Component>
  </didl:Item>
</didl:DIDL>

```



Complex Object Formats & OAI-PMH

- Resource represented via XML wrapper => OAI-PMH <metadata>
- Uniform solution for simple & compound objects
- Unambiguous expression of locator of datastream
- Disambiguation between locators & identifiers
- OAI-PMH datestamp changes whenever the resource (datastreams, secondary information) changes
- OAI-PMH semantics apply: “about” containers, set membership



OAI-PMH based approach using Complex Object Format

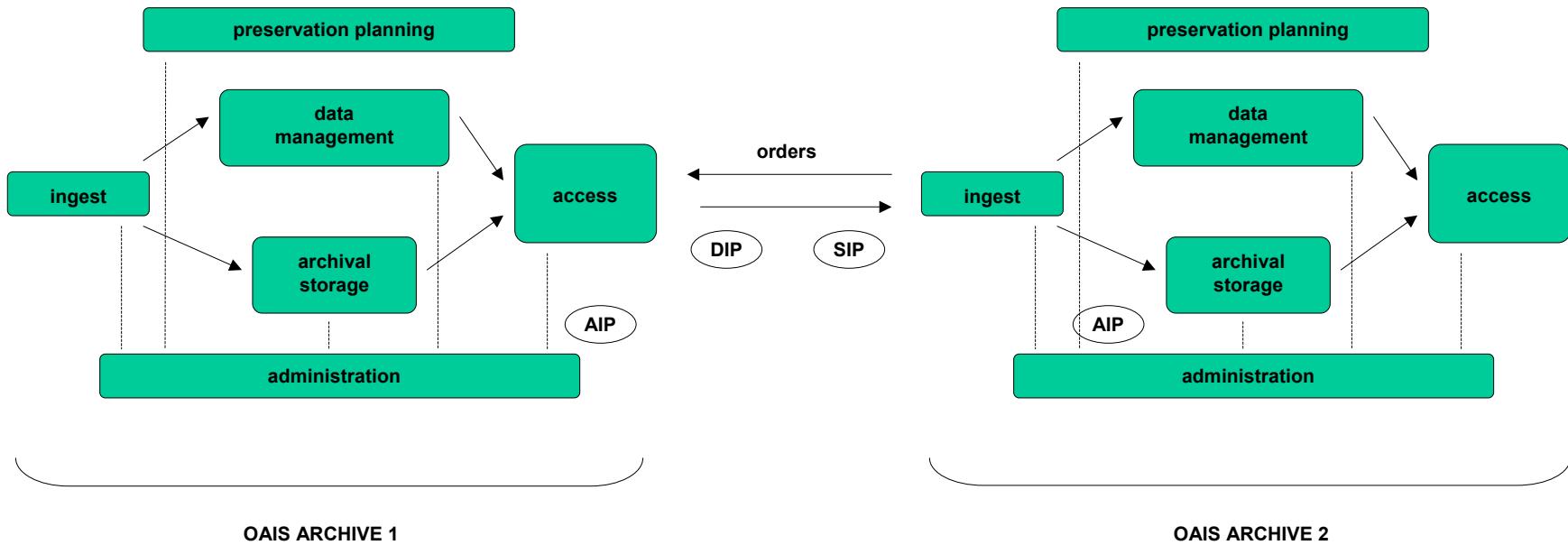
- Typical scenario:
 - an OAI-PMH harvester checks for support of a complex object format using the ListMetadataFormats verb
 - the harvester harvests the complex object metadata. Semantics of the OAI-PMH datestamp guarantee that new and modified resources are detected.
 - a parser at the end of the harvesting application analyzes each harvested complex object record:
 - The parser extracts the bitstreams that were delivered By-Value.
 - The parser extracts the unambiguous references to the network location of bitstreams delivered By-Reference.
 - a separate process, out-of-band from the OAI-PMH, collects the bitstreams delivered By-Reference from the extracted network locations.

Complex Object Formats & OAI-PMH : existing implementations

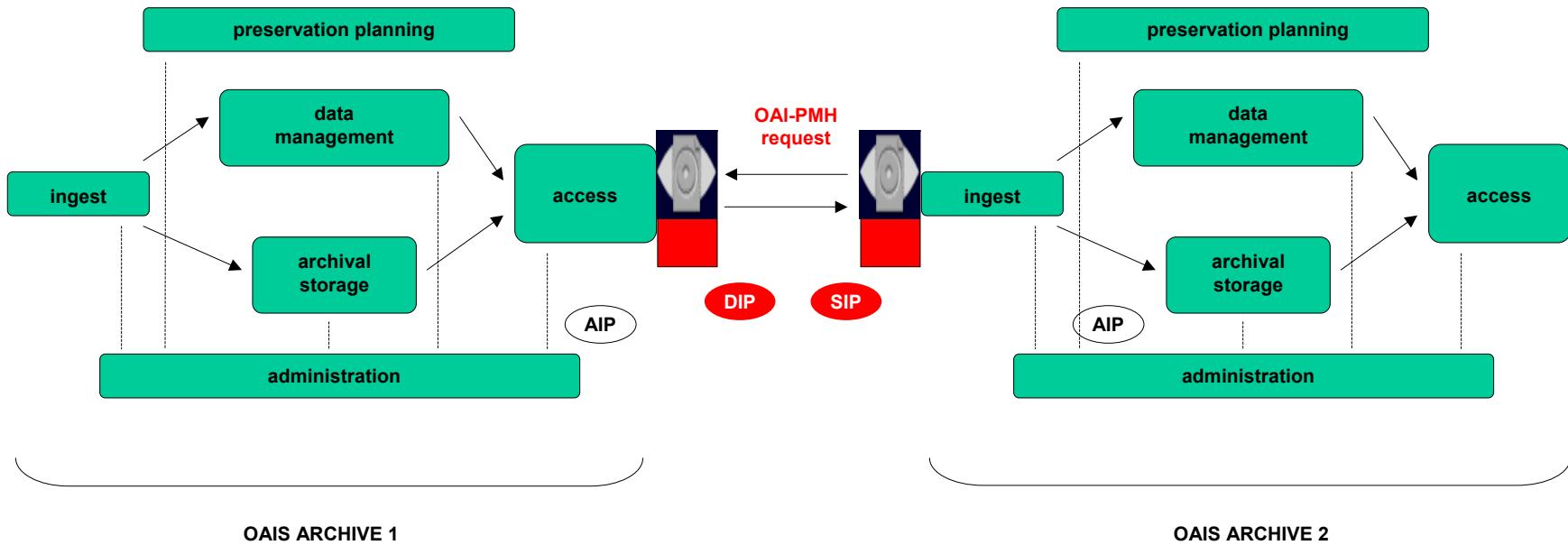


- aDORe: LANL repository
 - local storage of Terrabytes of scholarly assets
 - assets stored as MPEG-21 DIDL documents
 - DIDL documents made accessible to downstream applications via the OAI-PMH
- Mirroring of American Physical Society collection at LANL
 - maps APS document model to MPEG-21 DIDL Transfer Profile
 - exposes MPEG-21 DIDL documents through OAI-PMH infrastructure
 - includes W3C XML Signatures
- DSpace & Fedora plug-ins
 - maps DSpace/Fedora document model to MPEG-21 DIDL Transfer Profile
 - exposes MPEG-21 DIDL documents through OAI-PMH infrastructure
- mod_oai
 - Exposes resources on apache server through OAI-PMH
 - Introduces date-stamp bases harvesting to web servers

Complex Object Formats & OAI-PMH : archive export/ingest



Complex Object Formats & OAI-PMH : archive export/ingest



Complex Object Formats & OAI-PMH : issues

- Large records
- Making resources re-harvestable



Using Standards in Digital Library Design and Development
Jeroen Bekaert, Xiaoming Liu & Herbert Van de Sompel
JCDL 2005, June 7, 2005, Denver, CO





Software & readings

- Software:
 - DSpace DIDL plug-in (LANL), <http://didl-plug-in.sourceforge.net/>
 - modoai project (ODU, LANL), <http://www.modoai.org>
 - aDORe OAI-PMH harvester and de-referencing tool (LANL), to be released.
- Readings:
 - A standards-based solution for the accurate transfer of digital assets. To be published in DLib Magazine (June 2005 issue).
<http://www.dlib.org/dlib/june05/bekaert/06bekaert.html>
 - Resource Harvesting within the OAI-PMH Framework.
<http://www.dlib.org/dlib/december04/vandesompel/12vandesompel.html>

Z39.88-2004

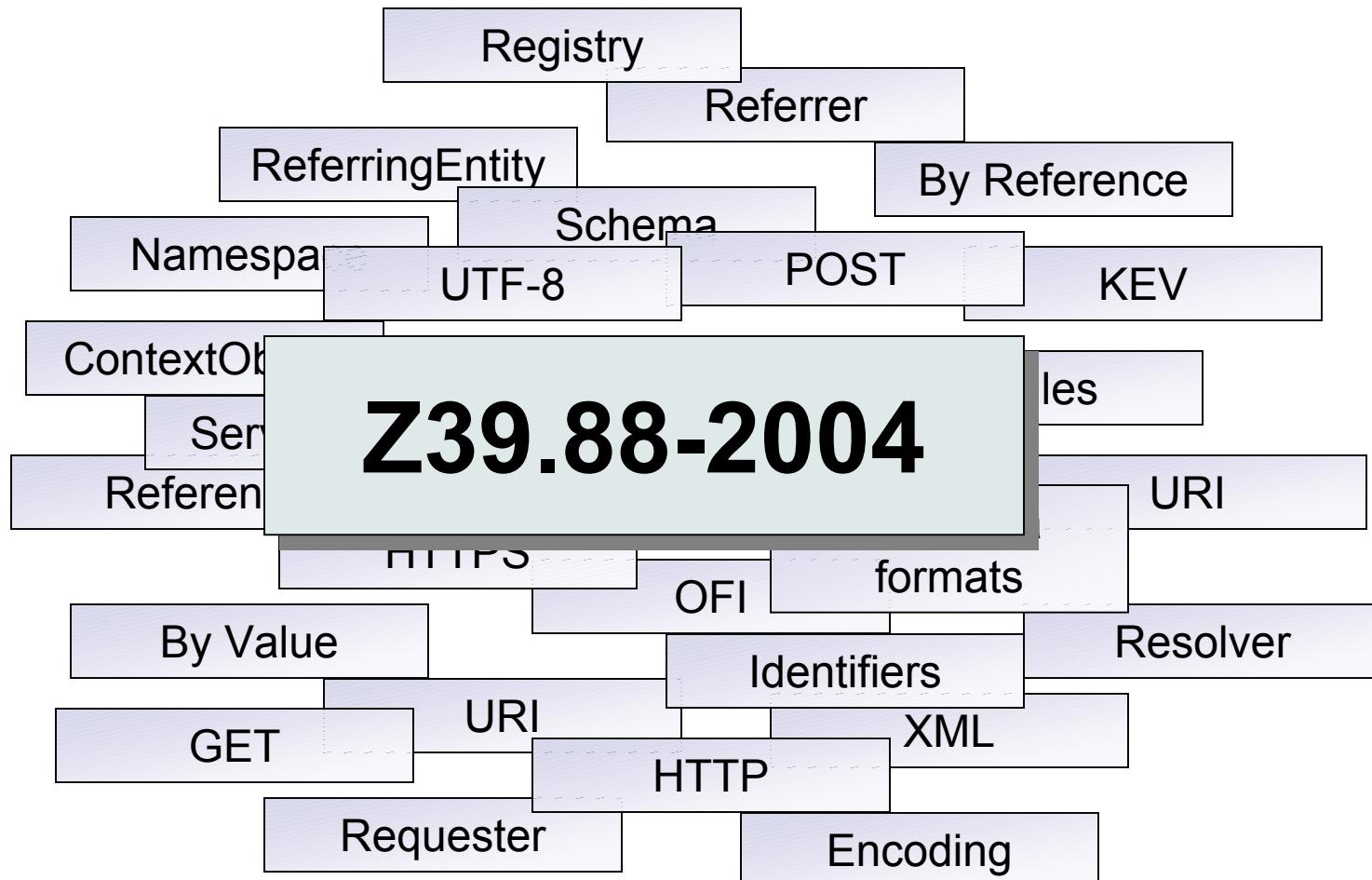
The OpenURL Framework

for

Context-Sensitive Services

Jeroen Bekaert, Xiaoming Liu, and Herbert Van de Sompel
Digital Library Research and Prototyping Team
Research Library, Los Alamos National Laboratory

Thanks for the nice slides:
Oliver Pesch
Chief Architect, EBSCO Publishing



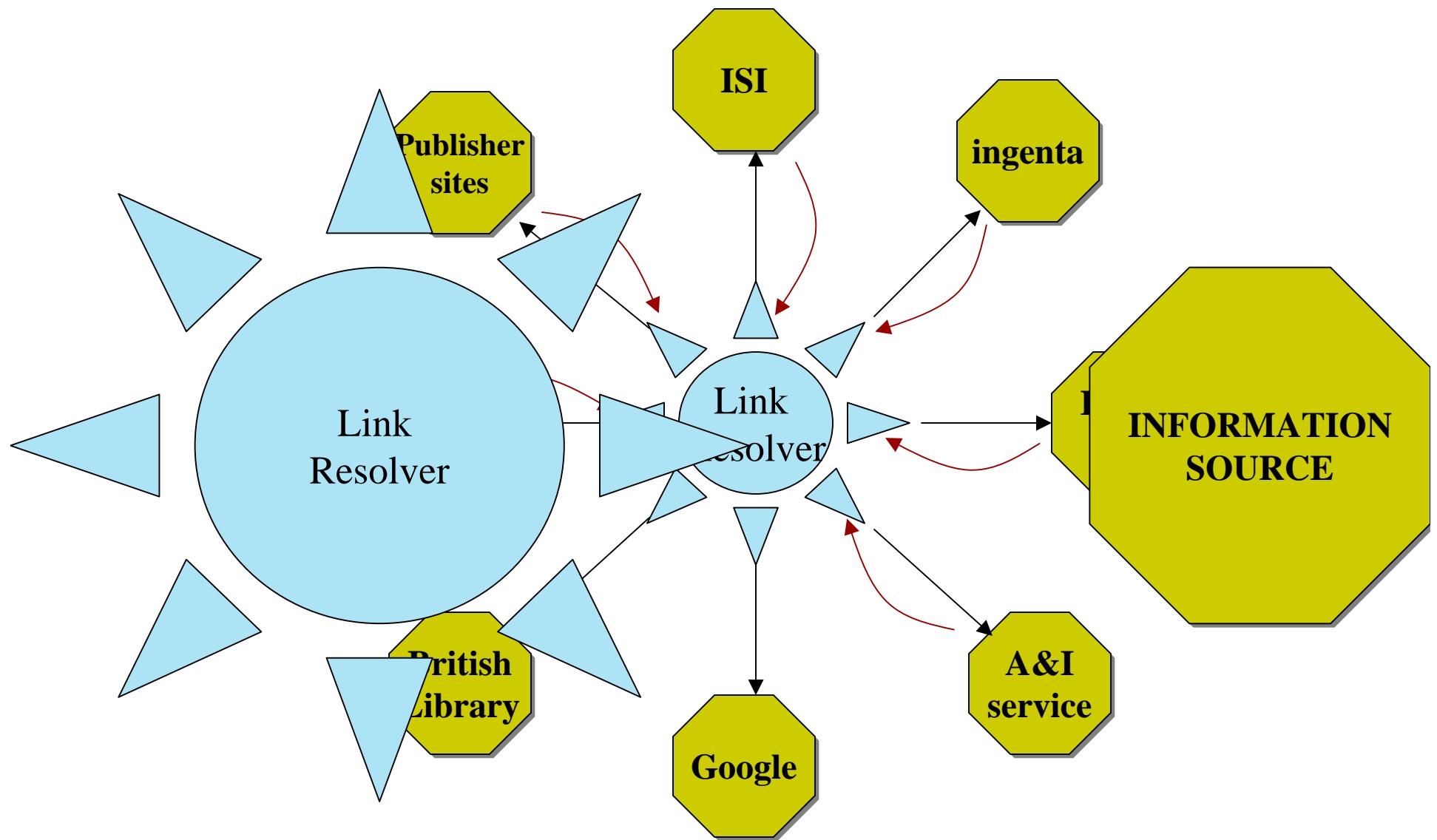
Topics

- What is a 0.1 OpenURL?
- Why the NISO OpenURL Standard?
- A tour of the NISO OpenURL Standard

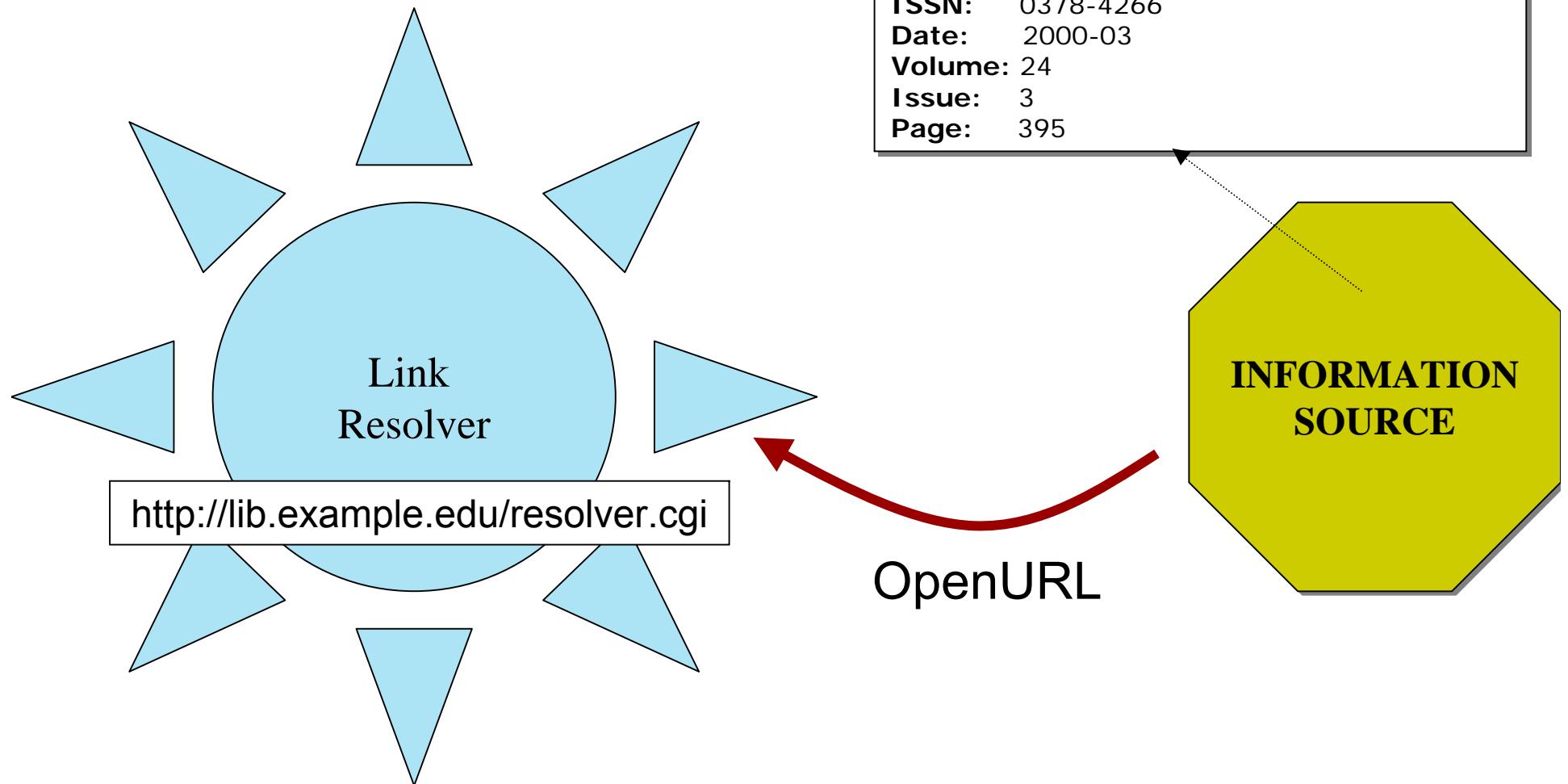
What is OpenURL 0.1 ?

- An accepted “standard” syntax for creating a link between an information source and a link resolver
- Pre-defines sets of data elements to use in describing an “item”
- Relies on HTTP protocol for transmission
- The **concept** of context-sensitive linking implemented for a specific class of resources: (some) scholarly assets

OpenURL 0.1



OpenURL 0.1



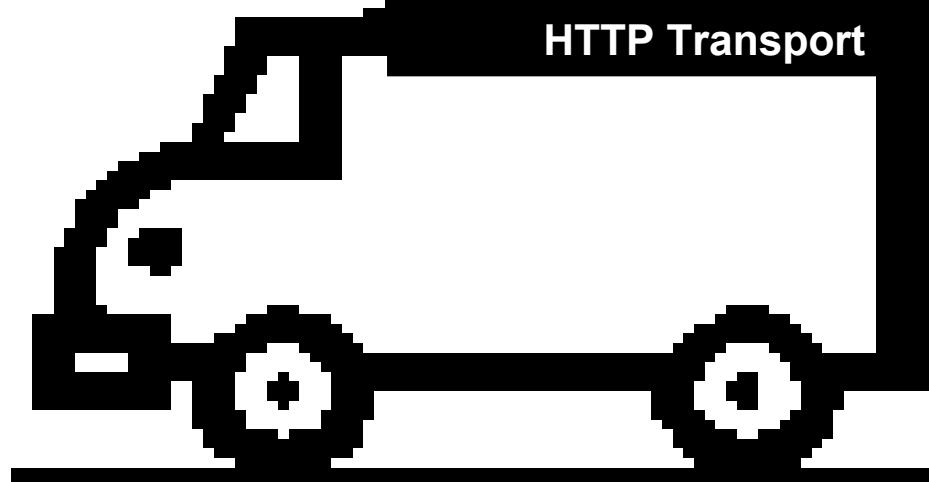
OpenURL 0.1

genre=article&
title=Journal of Banking and Finance&
issn=0378-4266&
date=2000-03&
volume=24&
issue=3&
spage=395&
aulast=Narayanan&
aufirst=Ranga&
atitle=Insider Trading and the Voluntary
Disclosure of Information by Firms&
cid=InfoSource

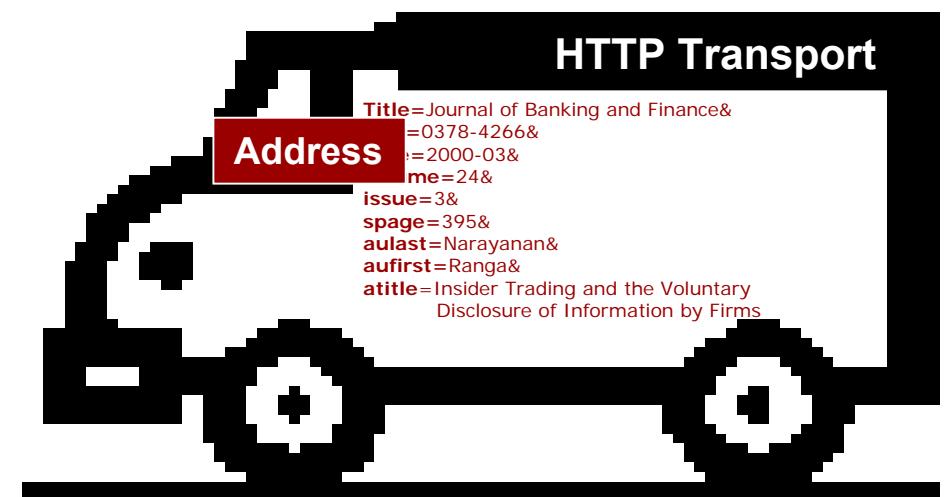
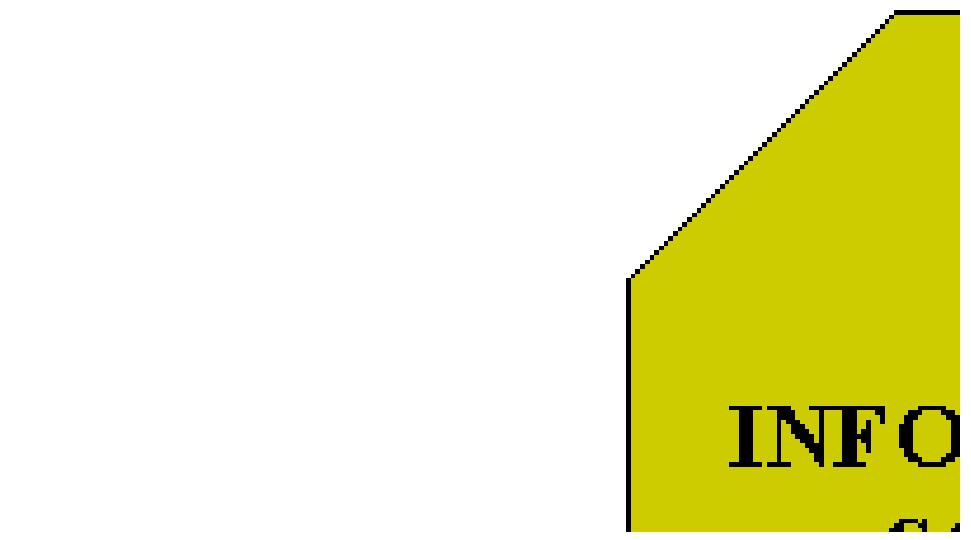
sid=InfoSource

<http://lib.exdu/resolver.cgi>

HTTP Transport



OpenURL 0.1



OpenURL 0.1 - limitations

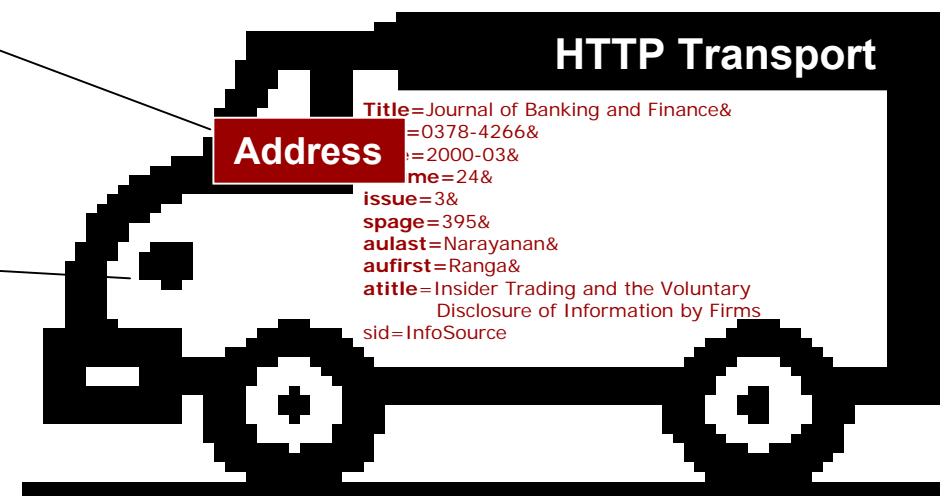
Allowable metadata genres and elements pre-defined with no means to define new ones

Only provides for key-value pair (HTTP GET or POST) representation of metadata.

Context of link limited to
-resolver (address)
-item
-source (sid)

OpenURL 0.1 is tied to HTTP transport

genre=article&
title=Journal of Banking and Finance&
issn=0378-4266&
date=2000-03&
volume=24&
issue=3&
spage=395&
aulast=Narayanan&
aufirst=Ranga&
atitle=Insider Trading and the Voluntary Disclosure of Information by Firms&
sid=InfoSource



Why the NISO OpenURL Standard?

- Ensure wide acceptance
- Facilitate emergence of Context-Sensitive Service Applications beyond the original OpenURL 0.1 community
- Address specific OpenURL 0.1 shortcomings
 - Support additional genres
 - Support richer data formats
 - Provide more complete context description
 - Allow ability to send request “by reference”
 - Support transports other than HTTP
- Provide an environment for OpenURL Applications to evolve in a controlled way

NISO OpenURL Standard

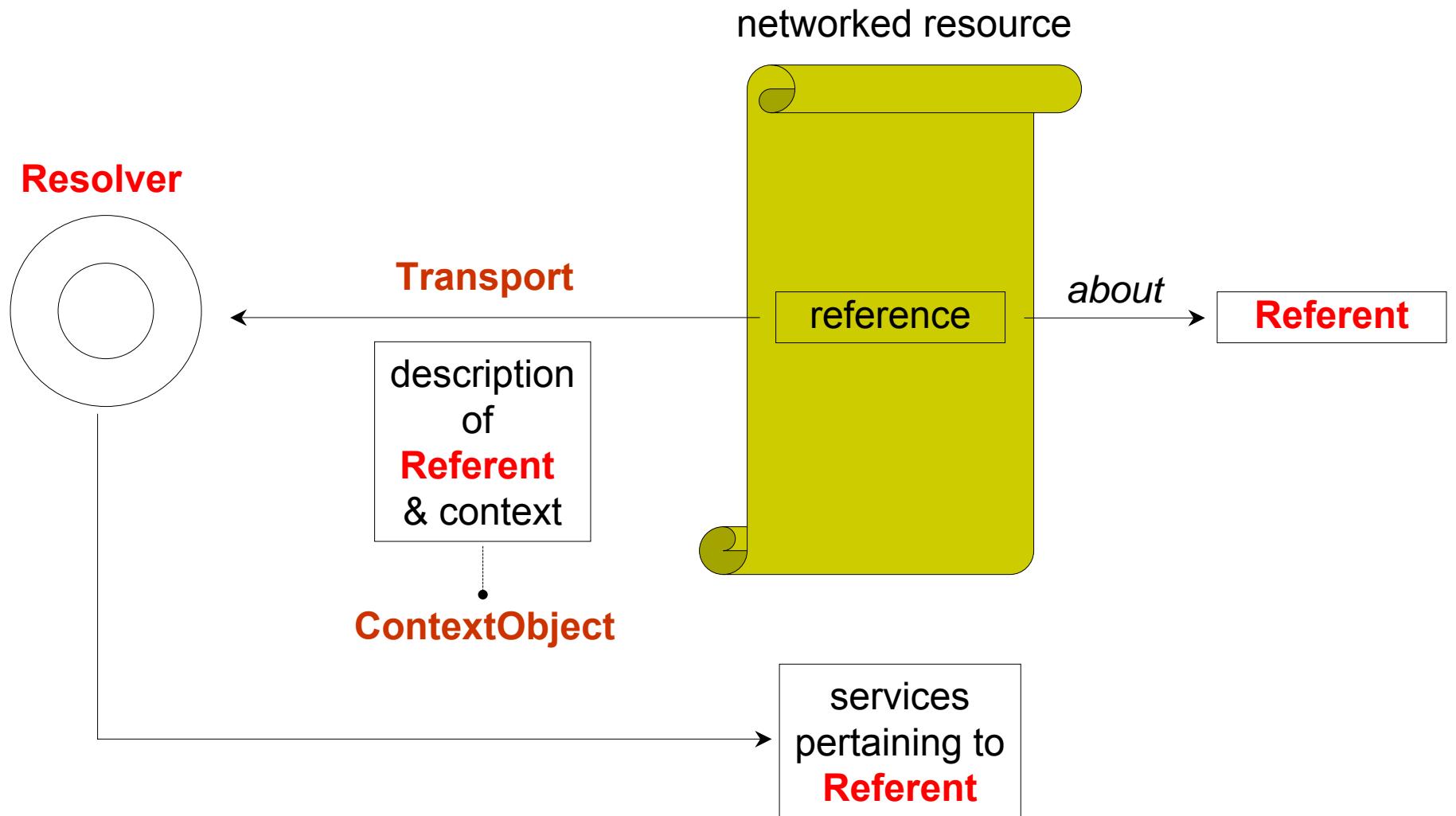
- A very generic specification that allows to implement **OpenURL Applications**
- **OpenURL Applications**: networked applications that implement the **concept** of context-sensitive services for a certain class of resources
- Based on generalization of original OpenURL ideas in D-Lib Bison-Fute paper

NISO OpenURL Standard

- Core Concept 1: The **ContextObject**
 - An “information package” that describes a referenced resource and the context within which it is being referenced
 - ContextObject has abstract definition (data model).
 - The data model can be instantiated via different representations: KEV, XML, RDF, ...

NISO OpenURL Standard

- Core Concept 2: **Transport of a ContextObject**
 - The idea is that ContextObjects will be transported in OpenURL Applications
 - Reason of transportation of a ContextObject: probably the request of services pertaining to the referenced resource
 - Transport of ContextObject is decoupled from representation of ContextObject => Can transport ContextObjects over HTTP, HTTPS, SOAP, OAI-PMH, ...



Deliverables from Committee AX

- 4 part standard
 - Part 1: ContextObject & Transport
 - Part 2: KEV ContextObject Format
 - Part 3: XML ContextObject Format
 - Part 4: OpenURL – HTTP(s) based - Transports
- Registry
- Community profiles: SAP-1 , SAP-2
- Implementation guidelines

Part 1: ContextObject and Transports

- Defines the general framework for specifying OpenURL Applications
- Introduces the ContextObject data model
- Introduces what it takes to represent a ContextObject
- Introduces Transports
- Defines Community Profiles as a means to define OpenURL Applications
- Defines the OpenURL Registry

ContextObject

An information construct with descriptions of 6 **Entities**:

- **Referent** (the resource that is being referenced)
- **Entities** that make up the context in which the **Referent** is referenced:
 - **ReferringEntity** (the resource that references the **Referent**)
 - **Requester** (the agent initiating the transportation of the ContextObject)
 - **ServiceType** (the purpose of transportation)
 - **Resolver** (the target of transportation)
 - **Referrer** (the system providing the ContextObject)

ContextObject



reference

ContextObject

Entities of the ContextObject can be described by means of the 1-4 Descriptors:

- Identifiers ~ many Namespaces
- By-Value Metadata ~ many Metadata Formats
- By-Reference Metadata ~ many Metadata Formats
- Private Data

 [The NISO AX Committee for the OpenURL Registry for the OpenURL Framework - ANSI/NISO Z39.88-2004](#)

[Repository Identification](#) | [Registry Entries](#) | [Implementation Guidelines](#) | [NISO OpenURL Version 0.1](#) | [Community Profile XML Format](#)

Core:Namespaces

<u>info:ofi/</u>	Namespace reserved for OpenURL Framework Registry Identifiers
<u>info:ofi/enc:</u>	Namespace reserved for Registry Identifiers of Character Encodings
<u>info:ofi/fmt:</u>	Namespace reserved for the identification of ContextObject Formats, Metadata Formats, Physical Representations, and Constraint Languages
<u>info:ofi/nam:</u>	Namespace reserved for Registry Identifiers of Namespaces.
<u>info:ofi/nam:data:</u>	Namespace for "data" URI Scheme
<u>info:ofi/nam:ftp:</u>	Namespace for "ftp" URI Scheme
<u>info:ofi/nam:http:</u>	Namespace for "http" URI Scheme
<u>info:ofi/nam:https:</u>	Namespace for "https" URI Scheme
<u>info:ofi/nam:info:</u>	Namespace for "info" URI Scheme
<u>info:ofi/nam:info:bibcode:</u>	Namespace of Astrophysics Bibcodes
<u>info:ofi/nam:info:doi:</u>	Namespace of Digital Object Identifiers
<u>info:ofi/nam:info:hdl:</u>	Namespace for CNRI handles
<u>info:ofi/nam:info:lccn:</u>	Namespace of Library of Congress Control Numbers
<u>info:ofi/nam:info:oai:</u>	Namespace of OAI Identifiers
<u>info:ofi/nam:info:oclcnum:</u>	Namespace of identifiers assigned by OCLC to records in the WorldCat database
<u>info:ofi/nam:info:pmid:</u>	Namespace of PubMed Identifiers
<u>info:ofi/nam:info:sici:</u>	Namespace of SICI Codes
<u>info:ofi/nam:info:sid:</u>	Namespace for identifiers that follow the info:sid scheme
<u>info:ofi/nam:ldap:</u>	Namespace for "ldap" URI Scheme

Find: extension

Core:Metadata Formats - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites Media Address http://alcme.oclc.org/openurl/servlet/OAIHandler?verb=ListRecords&metadataPrefix=oai_dc&set=Core:Metadata Go

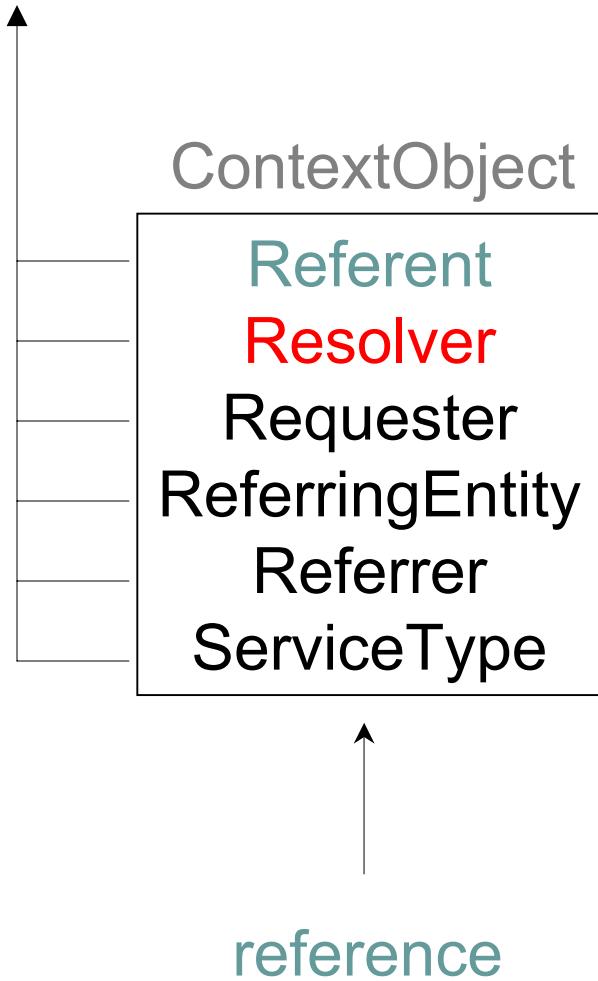
Registry for the OpenURL Framework - ANSI/NISO Z39.88-2003

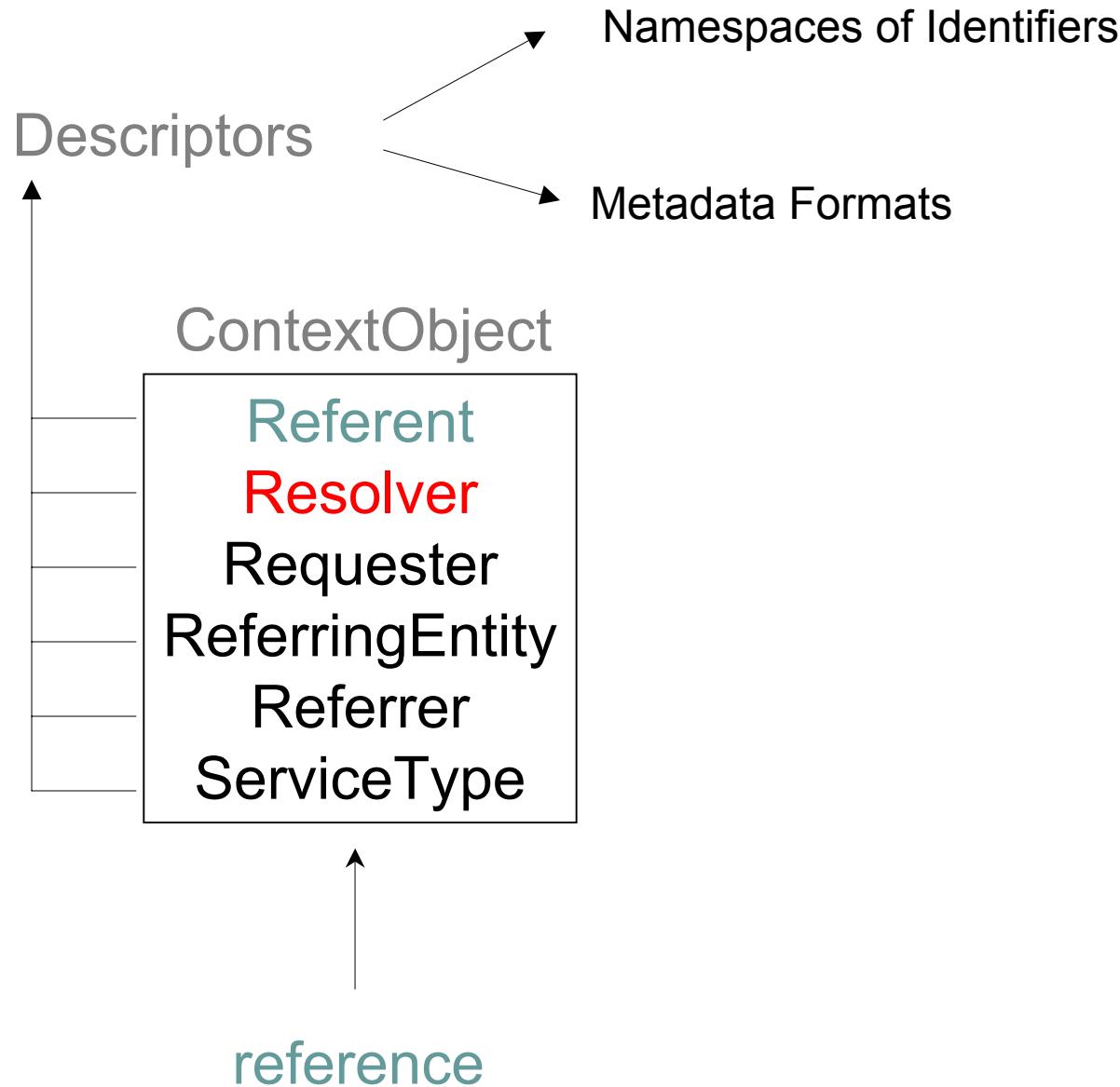
[Repository Identification](#) | [Registry Entries](#) | [Identifiers](#)

Core:Metadata Formats

<u>info:ofi/fmt:kev:mtx:book</u>	Key/Encoded-Value Metadata Format For Books
<u>info:ofi/fmt:kev:mtx:computer</u>	Draft Key/Encoded-Value Metadata Format For Computer Resources
<u>info:ofi/fmt:kev:mtx:dissertation</u>	Key/Encoded-Value Metadata Format For Dissertations
<u>info:ofi/fmt:kev:mtx:journal</u>	Key/Encoded-Value Metadata Format For Journals
<u>info:ofi/fmt:kev:mtx:map</u>	Draft Key/Encoded-Value Metadata Format For Maps
<u>info:ofi/fmt:kev:mtx:patent</u>	Key/Encoded-Value Metadata Format For Patents
<u>info:ofi/fmt:kev:mtx:score</u>	DRAFT Key/Encoded-Value Metadata Format For Music Scores
<u>info:ofi/fmt:kev:mtx:sound</u>	DRAFT Key/Encoded-Value Metadata Format For Sound Recordings
<u>info:ofi/fmt:kev:mtx:visual</u>	DRAFT Key/Encoded-Value Metadata Format For Visual Materials
<u>info:ofi/fmt:xml:xsd:book</u>	XML Metadata Format for Books
<u>info:ofi/fmt:xml:xsd:dissertation</u>	XML Metadata Format for Dissertations
<u>info:ofi/fmt:xml:xsd:journal</u>	XML Metadata Format for Journals
<u>info:ofi/fmt:xml:xsd:MARC21</u>	Library of Congress MARC XML Metadata Format
<u>info:ofi/fmt:xml:xsd:oai_dc</u>	Open Archives Initiative Unqualified Dublin Core
<u>info:ofi/fmt:xml:xsd:patent</u>	XML Metadata Format for Patents
<u>info:ofi/fmt:xml:xsd:pro</u>	XML Format to represent Community Profiles.

Descriptors





ContextObject Format

Standard introduces interesting formalization of a Format
as a triple consisting of choice for:

- Serialization: i.e. KEV, XML
- Constraint Language: i.e. Z39.88-2004 Matrix, XML Schema Language
- Constraint Definition: i.e. an XML Schema created to convey book metadata

ContextObject Format

The triple-formalism is used for:

- Metadata Formats: to describe Entities
- ContextObject Format: to represent ContextObjects

The triple-formalism is revealed in Registry Identifiers for Metadata Formats & ContextObject Formats:

- info:ofi/fmt:**kev:mxt:book**
- info:ofi/fmt:**xml:xsd:ctx**

Core:ContextObject Formats - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Favorites Media Address http://alcme.oclc.org/openurl/servlet/OAIHandler?verb>ListRecords&metadataPrefix=oai_dc&set=Core:Context Go

The NISO AX Committee for the OpenURL

Registry for the OpenURL Framework - ANSI/NISO Z39.88-2003

[Repository Identification](#) | [Registry Entries](#) | [Identifiers](#)

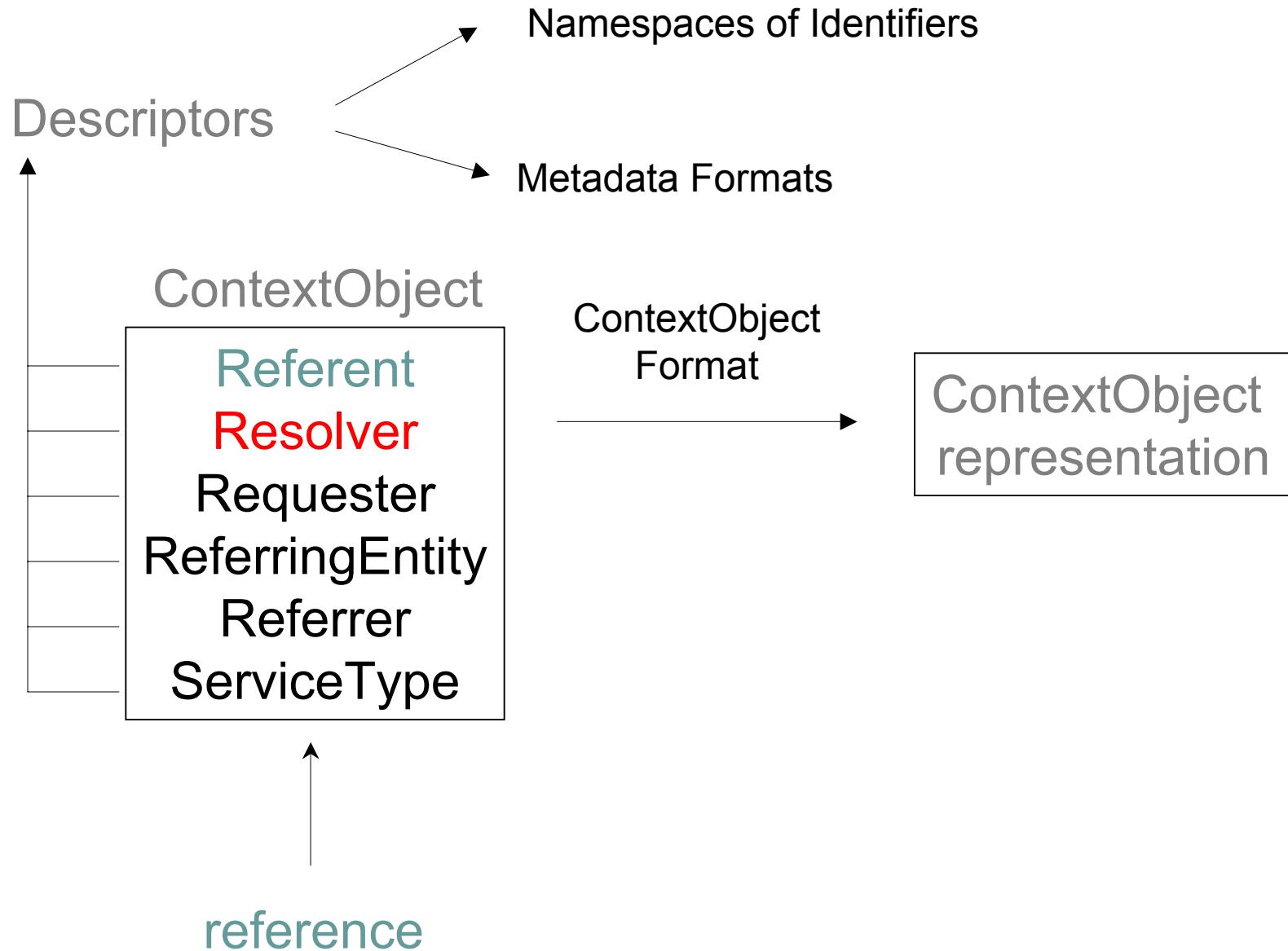
Core:ContextObject Formats

[info:ofi/fmt:kev:mtx:ctx](#) Key/Encoded-Value ContextObject Format
[info:ofi/fmt:xml:xsd:ctx](#) The XML ContextObject Format

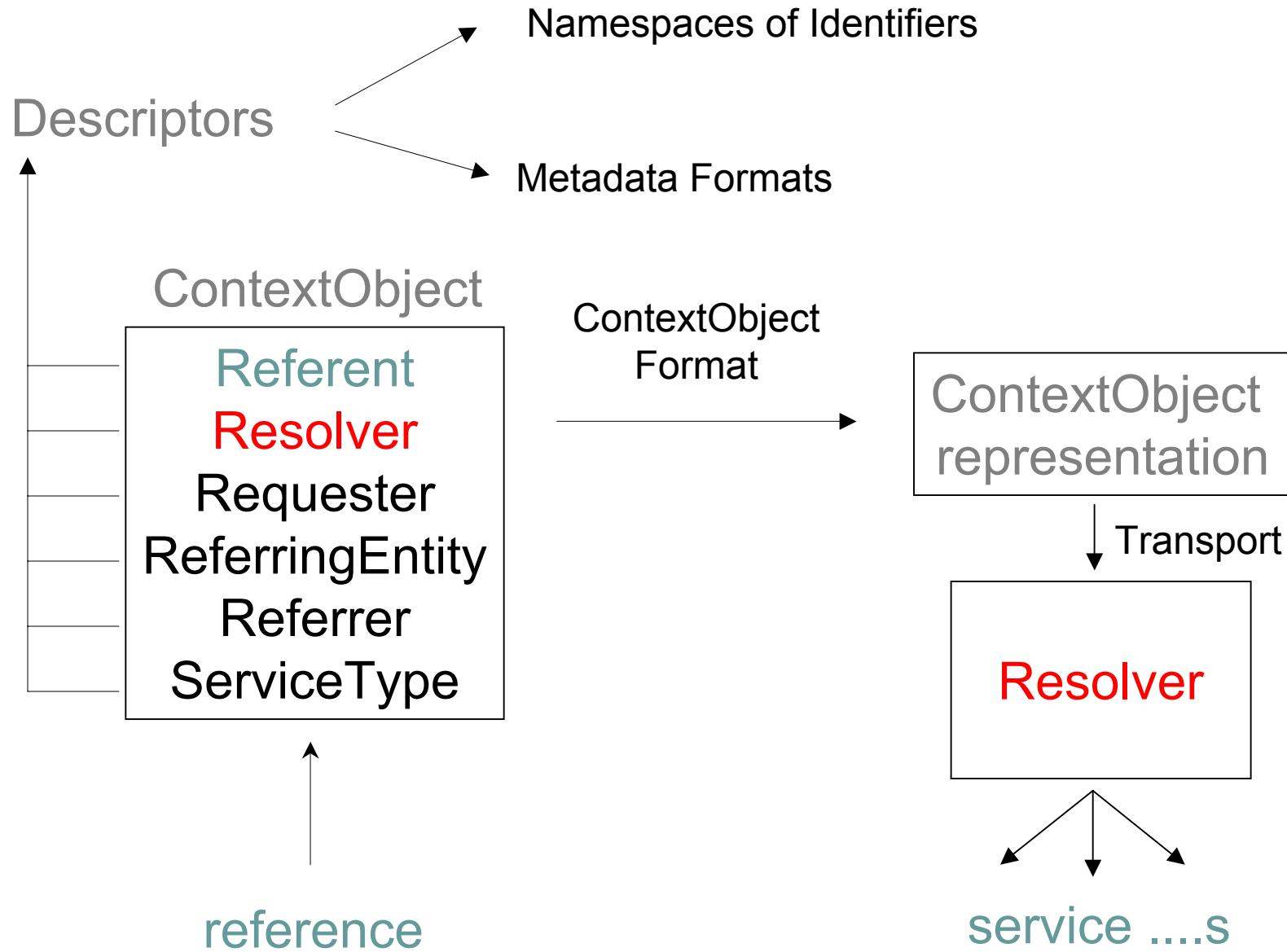
openurl.comment@library.caltech.edu

Internet



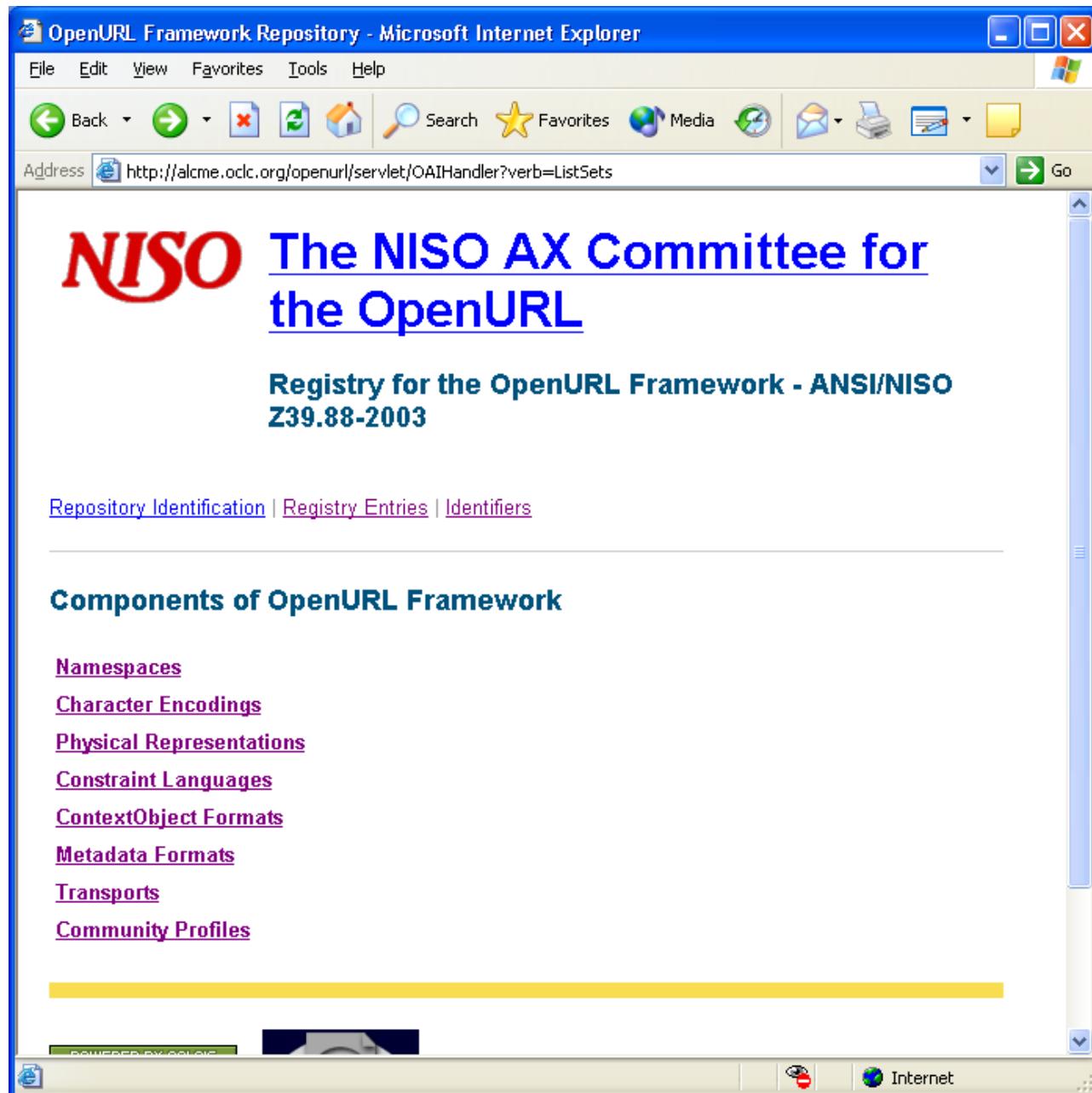
Transports

- A representation of a ContextObject can be **transported** in different ways, e.g.
 - HTTP(S) GET/POST
 - SOAP
 - OAI-PMH
 - ...



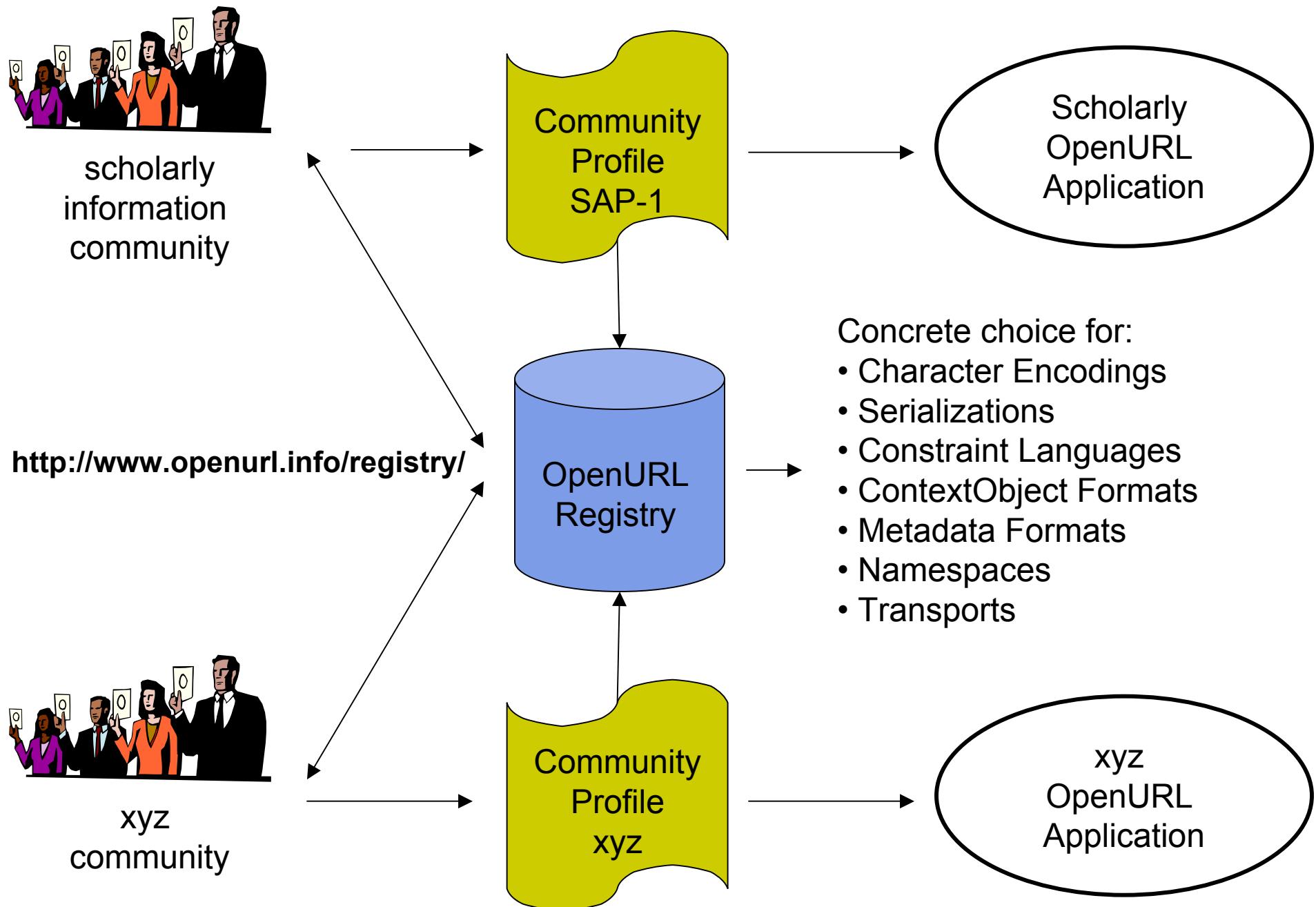
Registry

- At <http://www.openurl.info/registry/>
- Contains entries for all choices of the core components of the OpenURL Framework
- Registry comes pre-loaded to facilitate an OpenURL Application similar to OpenURL 0.1
- New entries can be registered
- Entries have Registry Identifiers in info:ofi/ namespace



Community Profiles

- A Community Profile summarizes the choices of core components of the OpenURL Framework for the creation of a specific OpenURL Application
- Machine readable, format defined by XML Schema
- Currently in Registry: `info:ofi/pro:sap-1` , `info:ofi/pro:sap-2`



Part 2: KEV ContextObject Format

- Represents a ContextObject as a list of ampersand-delimited Key/Encoded-Value pairs
- Similar to “payload” of OpenURL 0.1
- But extensible
- Format triple is (kev,mtx,ctx)
- Illustrated here for use in OpenURL Application similar to OpenURL 0.1

ContextObject Format Matrix - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Favorites Media Address http://alcme.oclc.org/openurl/servlet/OAIHandler/extension?verb=GetMetadata&metadataPrefix=mtx&identifier=info:ofi/fmt:kev Go

The Matrix

Delim	Key	Equals	Value	Min	Max	Description
#	ctx_			0	1	Administration. As Admin is an optional field in a ContextObject, any of the keys with prefix ctx_ may be present
&	ctx_ver	=	Z39.88-2003	0	1	ContextObject version. This has a fixed value
&	ctx_enc	=	<data>	0	1	ContextObject encoding. The value for ctx_enc specifies the character encoding used in the ContextObject. Legitimate values are taken from the IANA list at http://www.iana.org/assignments/character-sets . The values to be used in the ContextObject are those listed next to Name or - if available - the values with an indication of 'preferred MIME name' in the IANA list. UTF-8 is the default value, representing UTF-8 encoded Unicode.
&	ctx_id	=	<data>	0	1	ContextObject Identifier
&	ctx_tim	=	<time>	0	1	ContextObject timestamp. YYYY-MM-DD or YYYY-MM-DDThh:mm:ssTZD
#	rft_			1	1	Referent. As Referent is a mandatory Entity in a ContextObject, at least one of the keys with prefix rft_ must be present
&	rft_id	=	<id>	0	*	Referent Identifier. Multiple instances of rft_id do not indicate multiple Referents, but rather multiple ways to identify a single Referent
						Identifier of By-Value Metadata Format for a Referent.

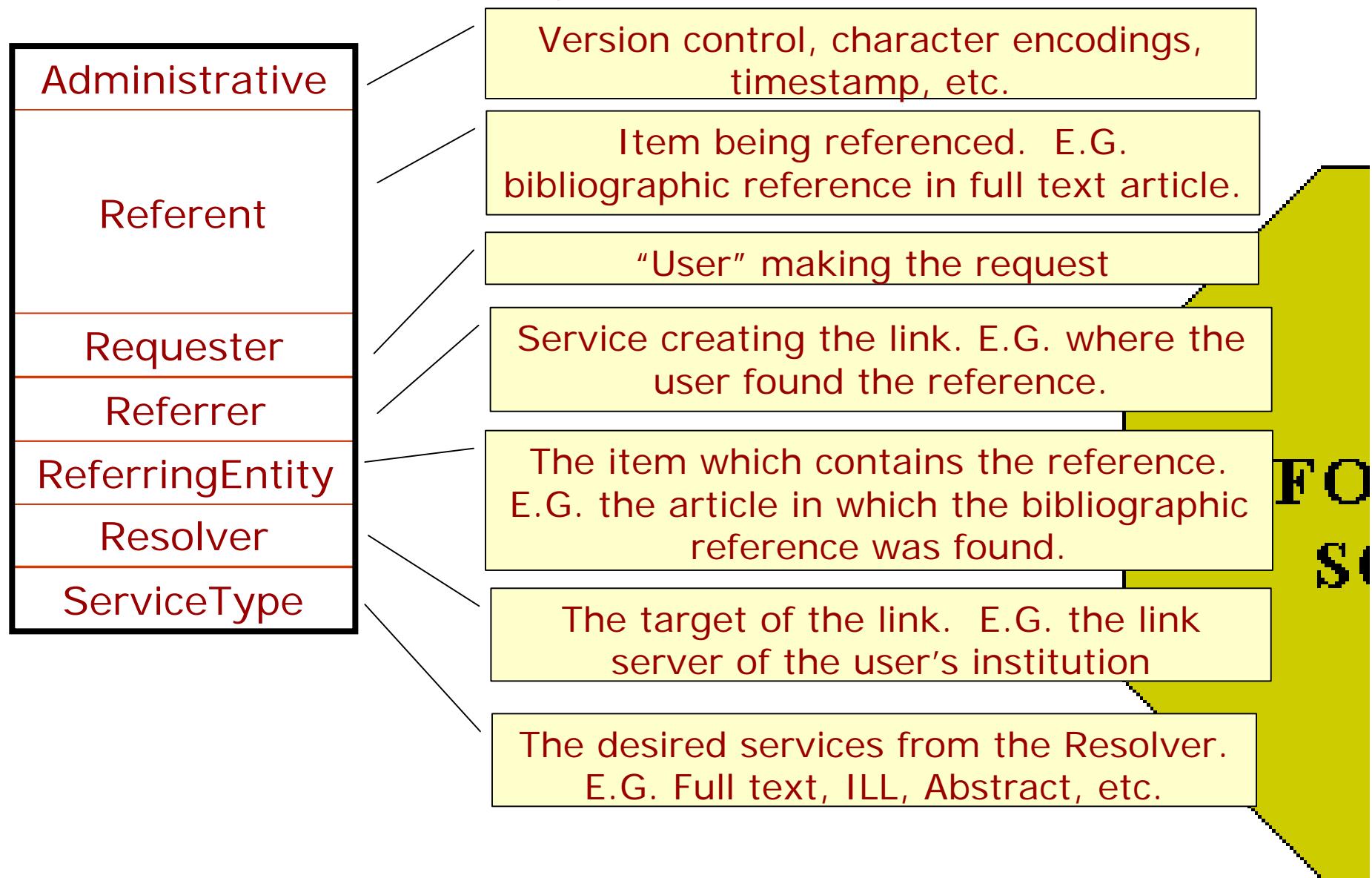
Done

ContextObject Format Matrix						
&	rft_id	=	<id>	0	*	indicate multiple Referents, but rather multiple ways to identify a single Referent
&	rft_val_fmt	=	<fmt-id>	0	1	Identifier of By-Value Metadata Format for a Referent. ORI, URI or XRI Identifier of the Metadata Format used for the description of the Referent through By-Value Metadata
#	rft_val	=		0	0	Reserved for future use
&	rft.<m-key>	=	<data>	0	*	By-Value Metadata Key for a Referent. The <m-key> is a Key defined in the KEV Metadata Format specified by the Value of the rft_val_fmt Key, which must be present. Use of the rft prefix is mandatory
&	rft_ref_fmt	=	<fmt-id>	0	1	By-Reference Metadata Format for a Referent. The rft_ref Key must also be present
&	rft_ref	=	<url>	0	1	By-Reference Metadata Location for a Referent. The rft_ref_fmt Key must also be present. The Resolver should retrieve the Metadata from the specified location
&	rft_dat	=	<data>	0	1	Referent Private Data
#	rfe_			0	1	ReferringEntity. As ReferringEntity is an optional Entity in a ContextObject, any of the keys with prefix rfe_ may be present
&	rfe_id	=	<id>	0	*	ReferringEntity Identifier. Multiple instances of rfe_id do not indicate multiple ReferringEntities, but rather multiple ways to identify a single ReferringEntity
&	rfe_val_fmt	=	<fmt-id>	0	1	Identifier of By-Value Metadata Format for a ReferringEntity. ORI, URI or XRI Identifier of the Metadata Format used for the description of the ReferringEntity through By-Value Metadata

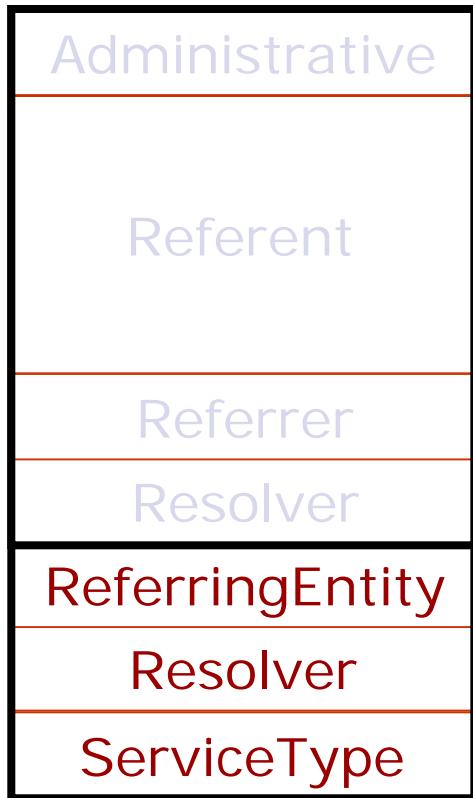
OpenURL 0.1 example

```
http://lib.example.edu/resolver.cgi?  
http://lib.example.edu/resolver.cgi?  
-----  
genre=article &  
title=Journal of Banking and Finance &  
issn=0378-4266 &  
date=2000-03 &  
volume=24 &  
issue=3 &  
spage=395 &  
aulast=Narayanan &  
aufirst=Ranga &  
atitle=Insider Trading and the Voluntary  
Disclosure of Information by Firms &  
sid=InfoSource
```

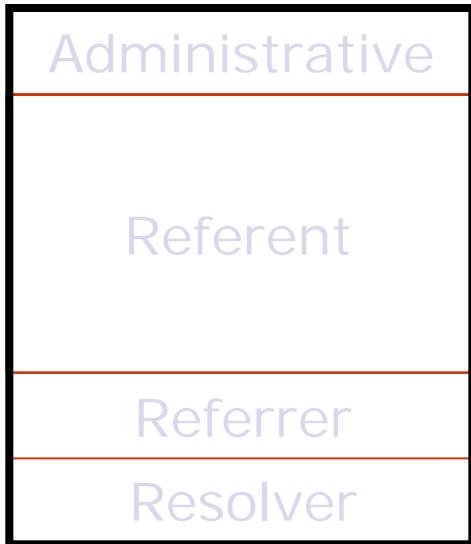
And the ContextObject is...



KEV ContextObject



KEV ContextObject

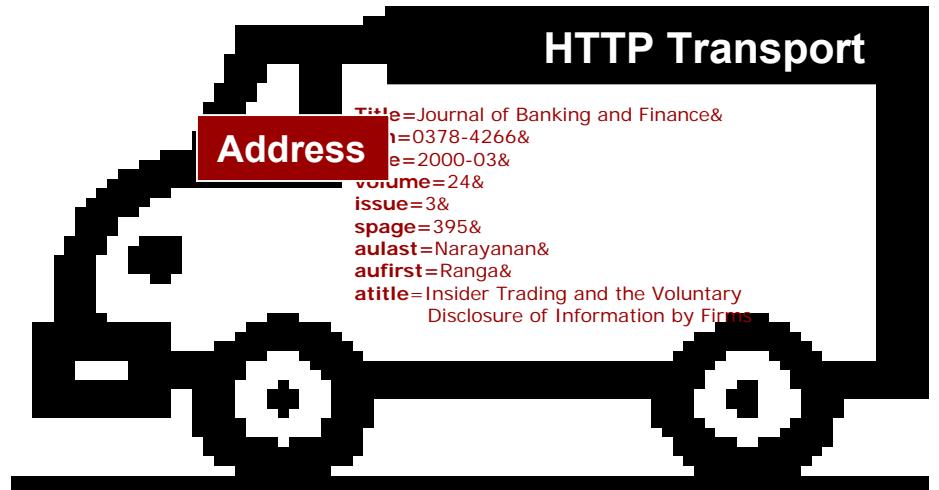


genre=article&
title=Journal of Banking and Finance&
issn=0378-4266&
date=2000-03&
volume=24&
issue=3&
spage=395&
aulast=Narayanan&
aufirst=Ranga&
atitle=Insider Trading and the Voluntary
Disclosure of Information by Firms&
rfr_id=info:sdid:InfoSource.com&

<http://lib.example.edu/resolver.cgi>

HTTP Transport

Title=Journal of Banking and Finance&
issn=0378-4266&
date=2000-03&
volume=24&
issue=3&
spage=395&
aulast=Narayanan&
aufirst=Ranga&
atitle=Insider Trading and the Voluntary
Disclosure of Information by Firms



KEV ContextObject



rft_val_fmt=info:ofi/fmt:kev:mtx:journal&
rtt.genre=article&
rft.title=Journal of Banking and Finance&
rft.issn=0378-4266&
rft.date=2000-03&
rft.volume=24&
rft.issue=3&
rft.spage=395&
rft.aulast=Narayanan&
rft.aufirst=Ranga&
rft.atitle=Insider Trading and the Voluntary Disclosure of Information by Firms&

res_id=http://lib.example.edu/resolver.cgi

KEV ContextObject

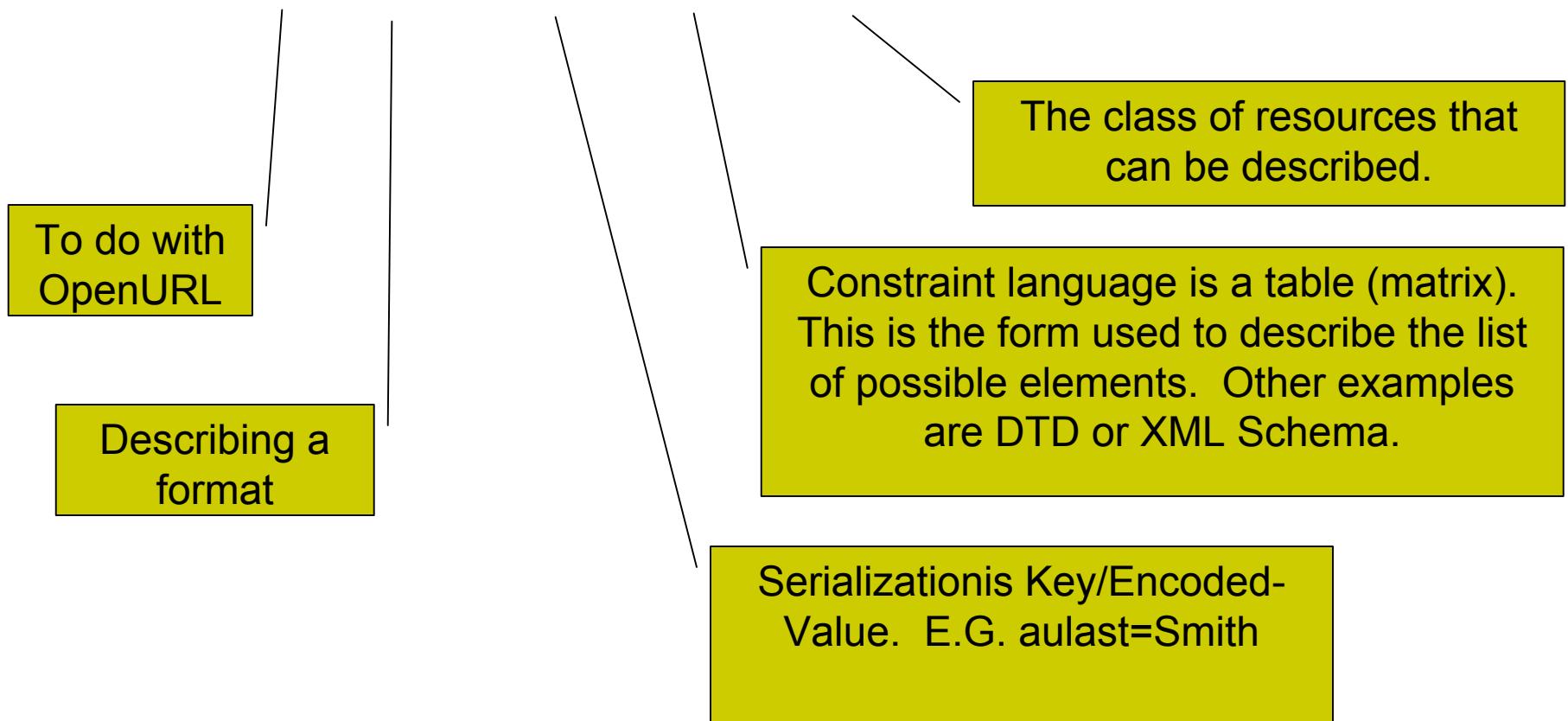


ctx_ver=Z39.88-2004&
ctx_ver=Z39.88-2004&
ctx_tim=2003-10-26&

INFO
SEARCH

Example of KEV Metadata Format

info:ofi/fmt:kev:mtx:journal



Core:Metadata Formats - ori:fmt:kev:mtx:journal - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Refresh Home Search Favorites Media Mail Print Find Links

Address http://alcme.oclc.org/openurl/servlet/OAIHandler?verb=GetRecord&metadataPrefix=mtx&identifier=ori:fmt:kev:mtx: Go Links

The Matrix

Delim	Key	Equals	Value	Min	Max	Description
&	aulast	=	<data>	0	1	First author's family name. This may be more than one word. In many citations, the author's family name is recorded first and is followed by a comma, i.e. Smith, Fred James is recorded as "aulast=smith"
&	aufirst	=	<data>	0	1	First author's given name or names or initials. This data element may contain multiple words and punctuation, i.e., "Fred James"
&	auinit	=	<data>	0	1	First author's first and middle initials.
&	auinit1	=	<data>	0	1	First author's first initial.
&	auinitm	=	<data>	0	1	First author's middle initial.
&	ausuffix	=	<data>	0	1	First author's name suffix. Qualifiers on an author's name such as "Jr.", "III" are entered here. i.e. Smith, Fred Jr. is recorded as "ausuffix=jr"
&	au	=	<data>	0	*	This data element contains the full name of a single author, i.e. "Smith, Fred M", "Harry S. Truman".
&	aucorp	=	<data>	0	1	Organization or corporation that is the author or creator of the document, i.e. "Mellon Foundation"
&	atitle	=	<data>	0	1	Article title.
&	title	=	<data>	0	1	Journal title. Provided for compatibility with version 0.1. Prefer jtitle.
&	jtitle	=	<data>	0	1	Journal title. Use the most complete title available. Abbreviated titles, when known, are records in stitle. This can also be expressed as title, for compatibility with version 0.1.

Part 3: XML ContextObject Format

- Represents a (list of) ContextObject(s) as an XML document
- Format triple is (xml,xsd,ctx)
- Far more expressive than KEV ContextObject Format

Part 4: OpenURL Transports

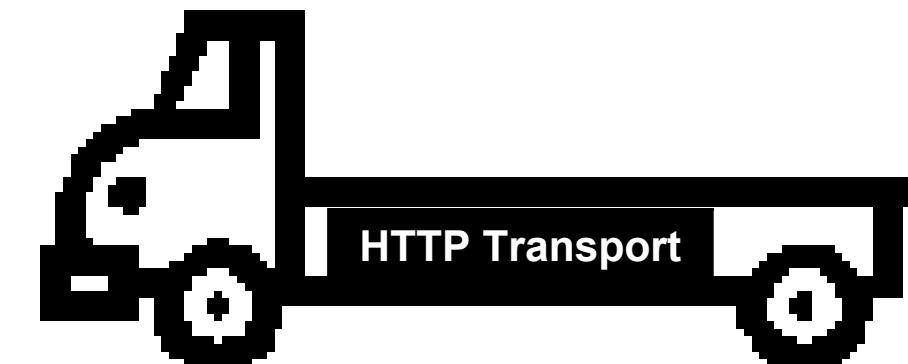
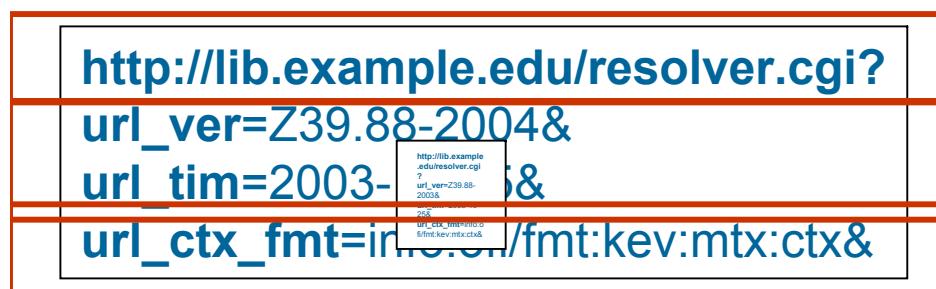
3 types of HTTP(S)-based manners to Transport ContextObjects:

- For all representations of ContextObjects:
 - By-Reference OpenURL
 - By-Value OpenURL
- For KEV ContextObjects only:
 - Inline OpenURL (very similar to OpenURL 0.1)

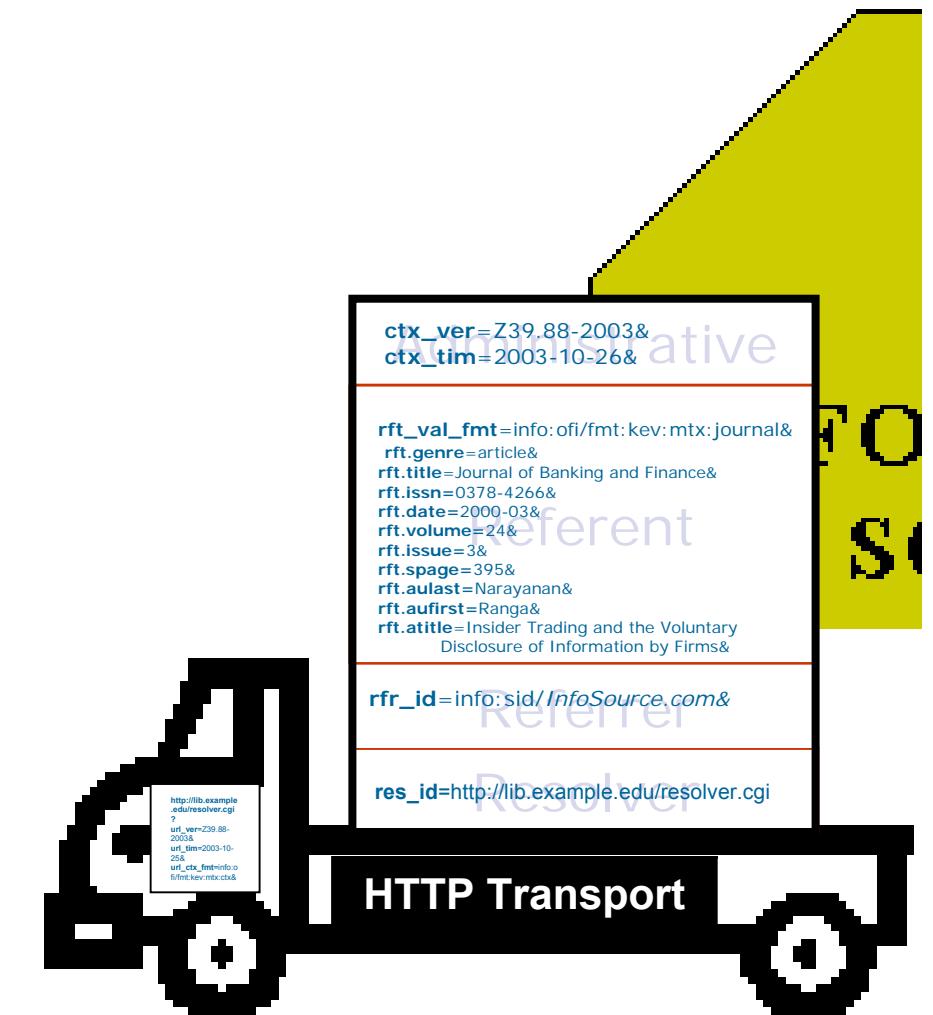
Flag on OpenURL identifies ContextObject Format:

- url_ctx_fmt=info:ofi/fmt:kev:mtx:ctx

KEV ContextObject, Inline OpenURL



KEV ContextObject, Inline OpenURL



Inline OpenURL

I am an
OpenURL

`http://example.org/myResolver`

`?url_ver=z39.88-2004`

`&url_ctx_tmt=info:ofi/fmt:kev:mtx:ctx`

ContextObject
Format

`&rft_val_fmt=info:ofi/fmt:kev:mtx:journal`

`&rtr_id=into:sid/myId.com:mydb`

`&rft_id=info:doi/10.1126/science.275.5304.1320`

`&rft_id=info:pmid/9036860`

`&rft.genre=article`

`&rft.atitle=Isolation of a common receptor for coxsackievirus B`

`&rft.title=Science`

`&rft.aulast=Bergelson`

`&rft.auinit=J`

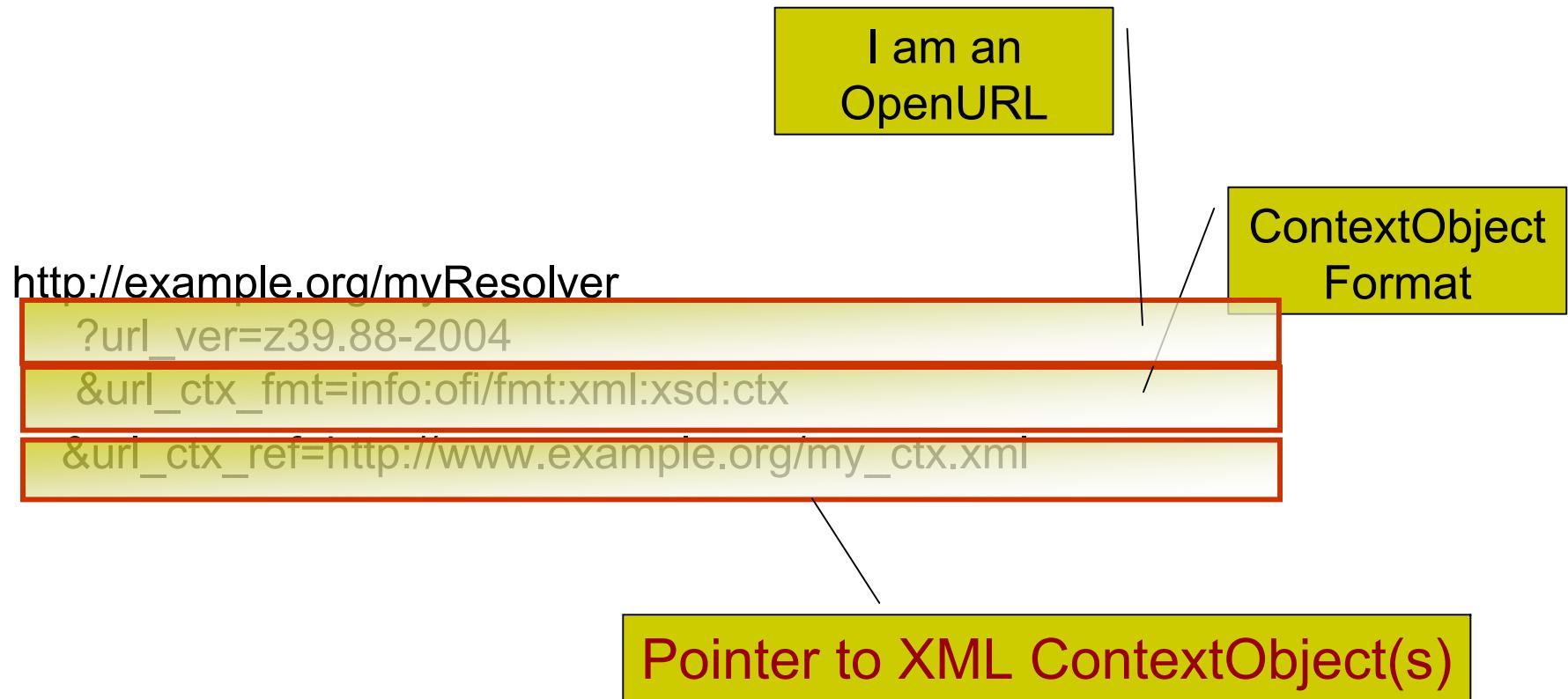
`&rft.date=1997`

`...`

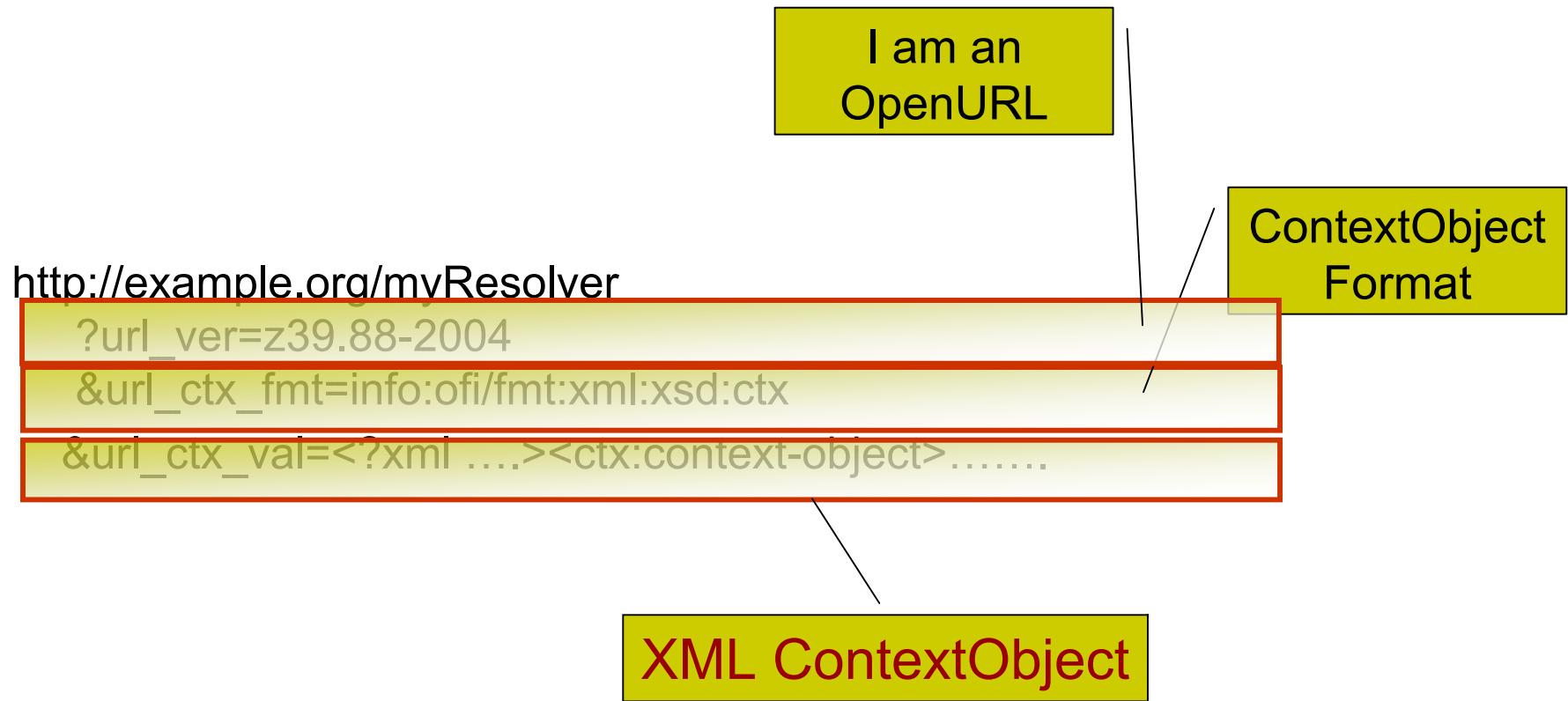
Metadata
Format

KEV ContextObject

By-Reference OpenURL



By-Value OpenURL



Would typically use this with HTTP(S) POST

Implementation Guidelines

- Centered on SAP-1
- For implementers
- Describe necessary bits of the standard
- How to create KEV OpenURLs
- Demonstrates the upgrade path from OpenURL 0.1 to OpenURL 1.0
- Includes hybrid OpenURLs

Using info URI to facilitate the referencing of information assets under the URI allocation

Jeroen Bekaert, Xiaoming Liu, and Herbert Van de Sompel
Digital Library Research & Prototyping Team
Research Library, Los Alamos National Laboratory



Using Standards in Digital Library Design and Development
Jeroen Bekaert, Xiaoming Liu & Herbert Van de Sompel
JCDL 2005, June 7, 2005, Denver, CO



info URI - motivation

- Original motivation
 - turn ‘legacy’ identifiers into URIs in the context of OpenURL Applications
 - Web-ify the identified resources
 - pmid 123456 => info:pmid/123456
- But also ‘new’ identifier namespaces
 - info:fedora/
 - info:lanl-repo/



info URI – what is?

- Low-barrier approach to URI-zing identifiers
- Offers registry of namespaces: <http://info-uri.info>
- Persistence of namespace identifiers guaranteed by operator of registry
- No claim re persistence of the identifiers (they are ‘inherited’)
- No expectation regarding resolution of identifiers: identification is at the core. But:
 - registered namespaces may have – local - resolution mechanisms
 - federated resolution systems may emerge

Mozilla Firefox	
File Edit View Go Bookmarks Tools Help	
http://info-uri.info/registry/OAIHandler?verb=ListRecords&metadataPrefix=oai_dc	
Getting Started Latest Headlines	
<u>info:bibcode/</u>	Namespace of Astrophysics Data System bibcodes
<u>info:ddbj-embl-genbank/</u>	Namespace of identifiers for sequence records in DDBJ/EMBL/GenBank
<u>info:doi/</u>	Namespace of Digital Object Identifiers
<u>info:fedora/</u>	Namespace of Fedora Digital Objects and Disseminations
<u>info:lanl-repo/</u>	Namespace of identifiers used in the Repository of the LANL Research Library
<u>info:lccn/</u>	Namespace of Library of Congress Control Numbers
<u>info:netref/</u>	Namespace of NISO Standard for Network Reference Services
<u>info:nla/</u>	Namespace of identifiers for the National Library of Australia's digital & digitised collections
<u>info:oclcnum/</u>	Namespace of OCLC Worldcat Control Numbers
<u>info:ofi/</u>	Namespace of Registry Identifiers used by the NISO OpenURL Framework Registry
<u>info:pmid/</u>	Namespace of identifiers of PubMed records
<u>info:refseq/</u>	Namespace of identifiers for RefSeq reference sequence record
<u>info:rlqid/</u>	Namespace of RLG Database Record identifiers
<u>info:sici/</u>	Namespace of Serial Item and Contribution Identifiers
<u>info:sid/</u>	Namespace of Source Identifiers used in the NISO OpenURL Framework
<u>info:srw/</u>	Namespace of Search/Retrieve Web Services
<u>info:ugent-repo/</u>	Namespace of identifiers for the University Library Ghent information assets



Using Standards in Digital Library Design and Development
 Jeroen Bekaert, Xiaoming Liu & Herbert Van de Sompel
 JCDL 2005, June 7, 2005, Denver, CO



Info URI – resolution?

- Decoupling of identification and resolution
 - identification remains stable over long term
 - actual resolution mechanism can change over time
- Stable interface to resolution via OpenURL?
 - http://my.openurl.org/resolve?rft_id=info:pmid/123456 &...
 - ContextObject Format, Transport, etc can evolve over time due to generic specification
- A level of indirection for the resolution system itself





readings

- The "info" URI Scheme for Information Assets with Identifiers in Public Namespaces: <http://www.ietf.org/internet-drafts/draft-vandesompel-info-uri-03.txt>
- info URI FAQ : <http://info-uri.info/registry/docs/misc/faq.html>
- Info URI Registry <http://info-uri.info>

aDORe: a modular and standards-based Digital Object repository

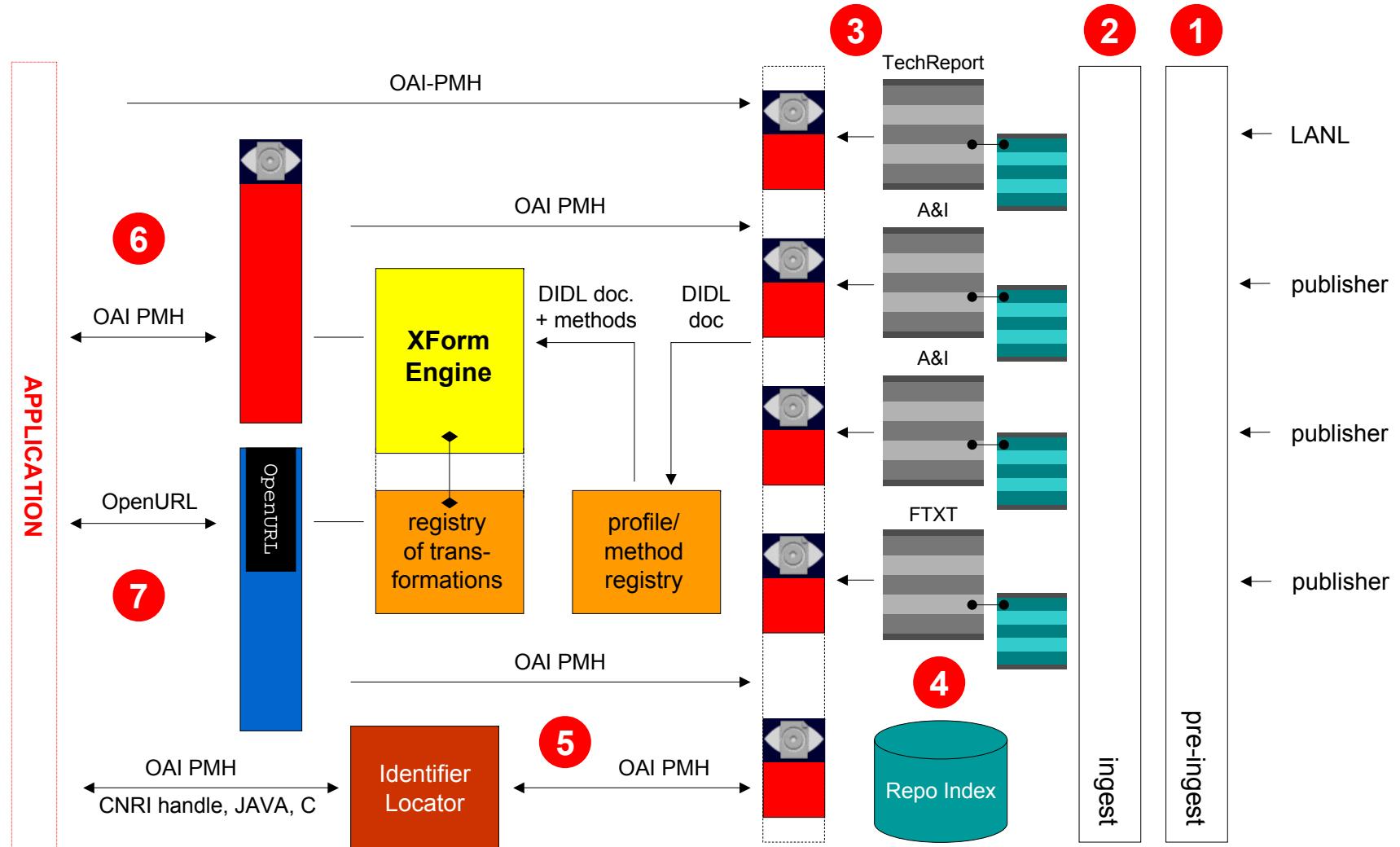
Jeroen Bekaert, Xiaoming Liu, and Herbert Van de Sompel
Digital Library Research & Prototyping Team
Research Library, Los Alamos National Laboratory



Using Standards in Digital Library Design and Development
Jeroen Bekaert, Xiaoming Liu & Herbert Van de Sompel
JCDL 2005, June 7, 2005, Denver, CO



overview of the aDORe architecture



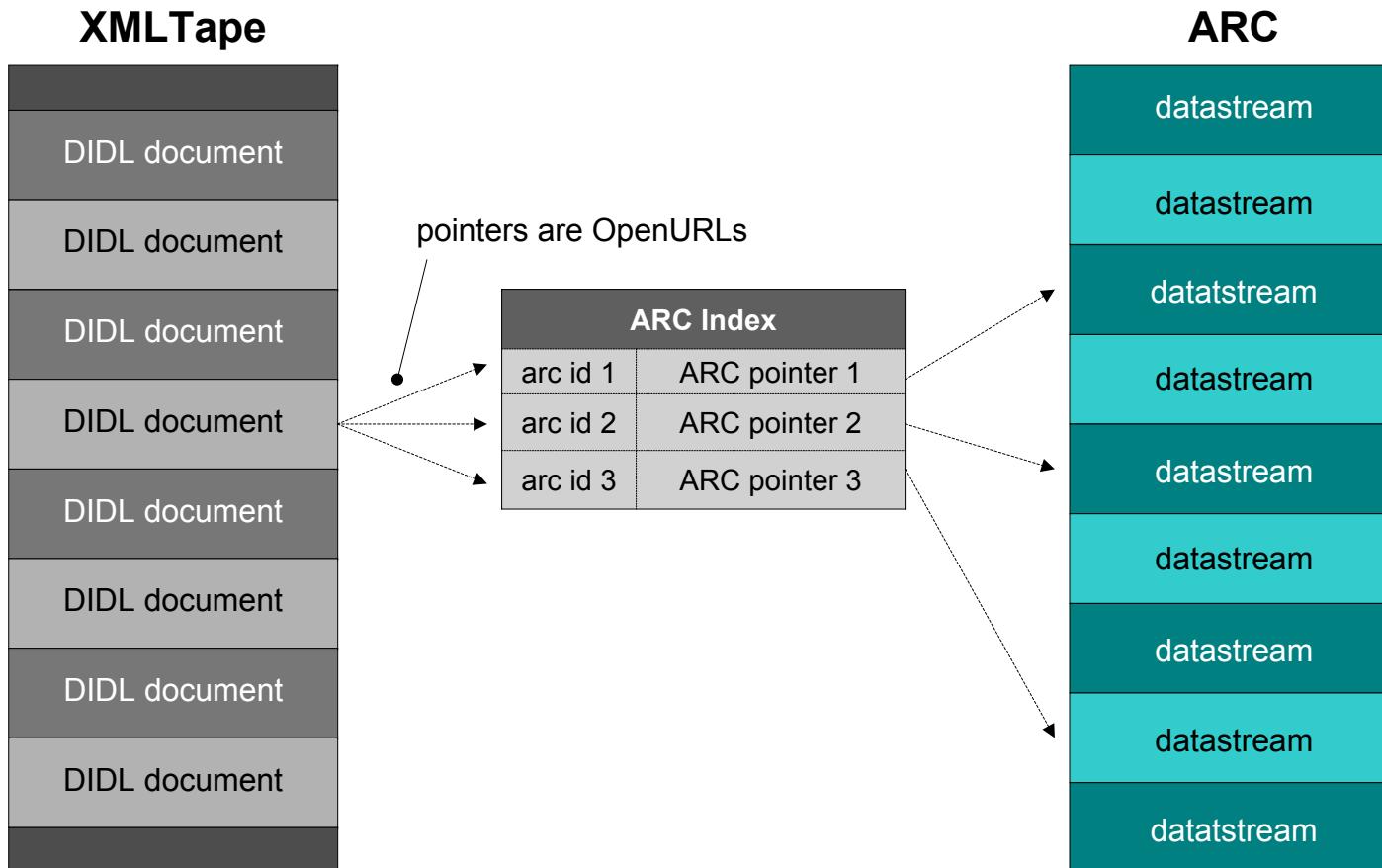
pre-ingest: data input from information provider

- Data feeds from third parties:
 - delivered in various ways (http, ftp, OAI-PMH, ..)
 - many different formats
 - typically contain many assets in a single feed
 - assets are typically ‘complex’, i.e. they consist of multiple datastreams

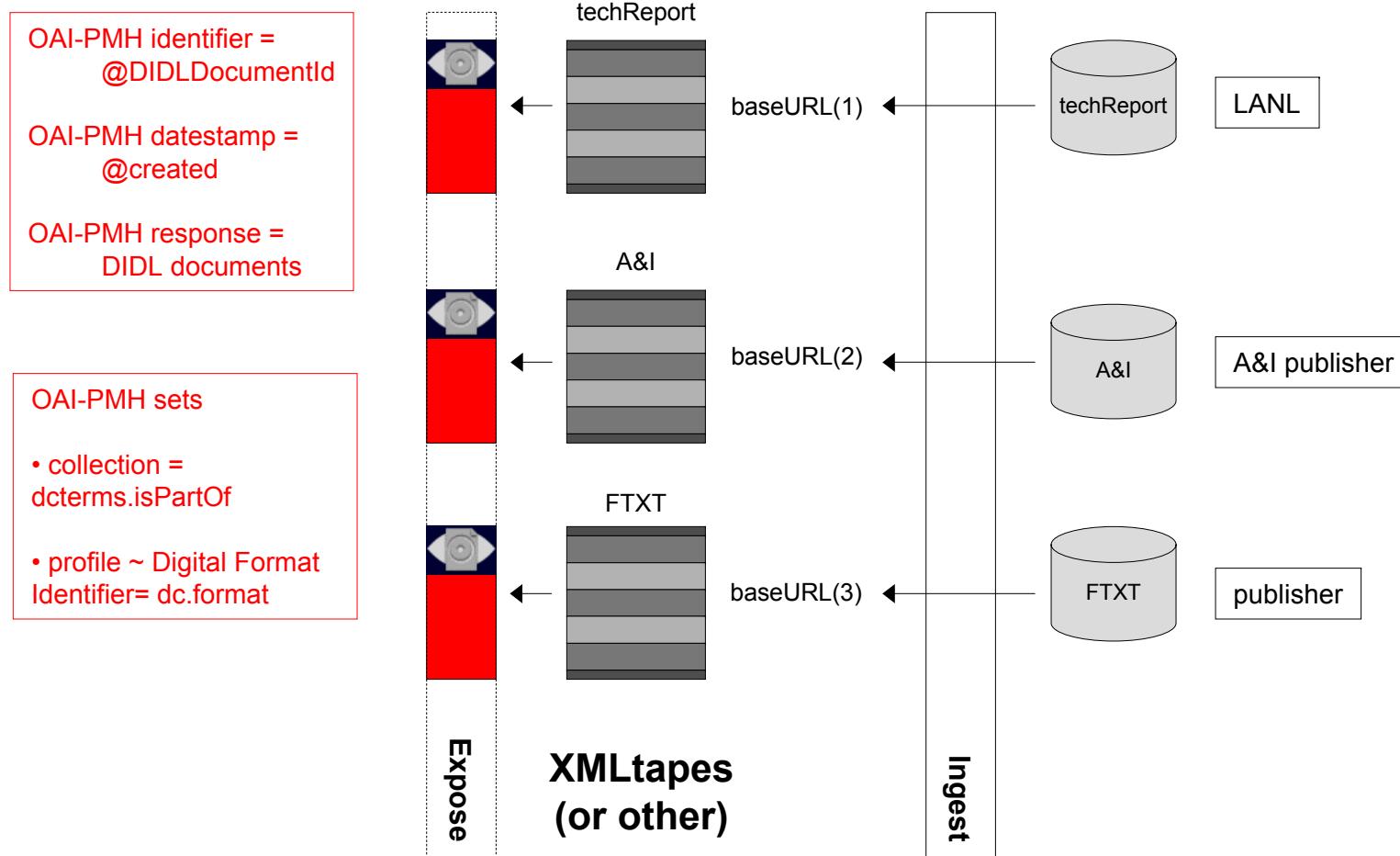
ingest: creation of DIDL documents

- New deliveries are processed for ingestion. For each delivered asset, a DIDL document is created:
 - a top-level `Item` element is introduced representing the asset
 - constituents of the asset are provided as child elements of this top-level `Item`.
 - `sub-Item` elements are used when the constituent of the asset has a Content Identifier
 - Component/Resource constructs are used when the constituent has no Content Identifier
 - the DIDL document contains inline XML data (such as MARCXML, ...)
 - the DIDL document contains pointers to bitstreams stored in ARC files
 - secondary information pertaining to the Digital Object is conveyed using Descriptor/Statement constructs
 - MPEG-21 DII is used to convey the Content Information Identifiers
 - secondary information pertaining to the DIDL document is conveyed using the `DIDLInfo` element
- DIDL documents are the OAIS AIPs in the aDORe repository
- DIDL documents become uniform proxies to the heterogeneous assets

XMLtapes for DIDs, ARC files for bitstreams

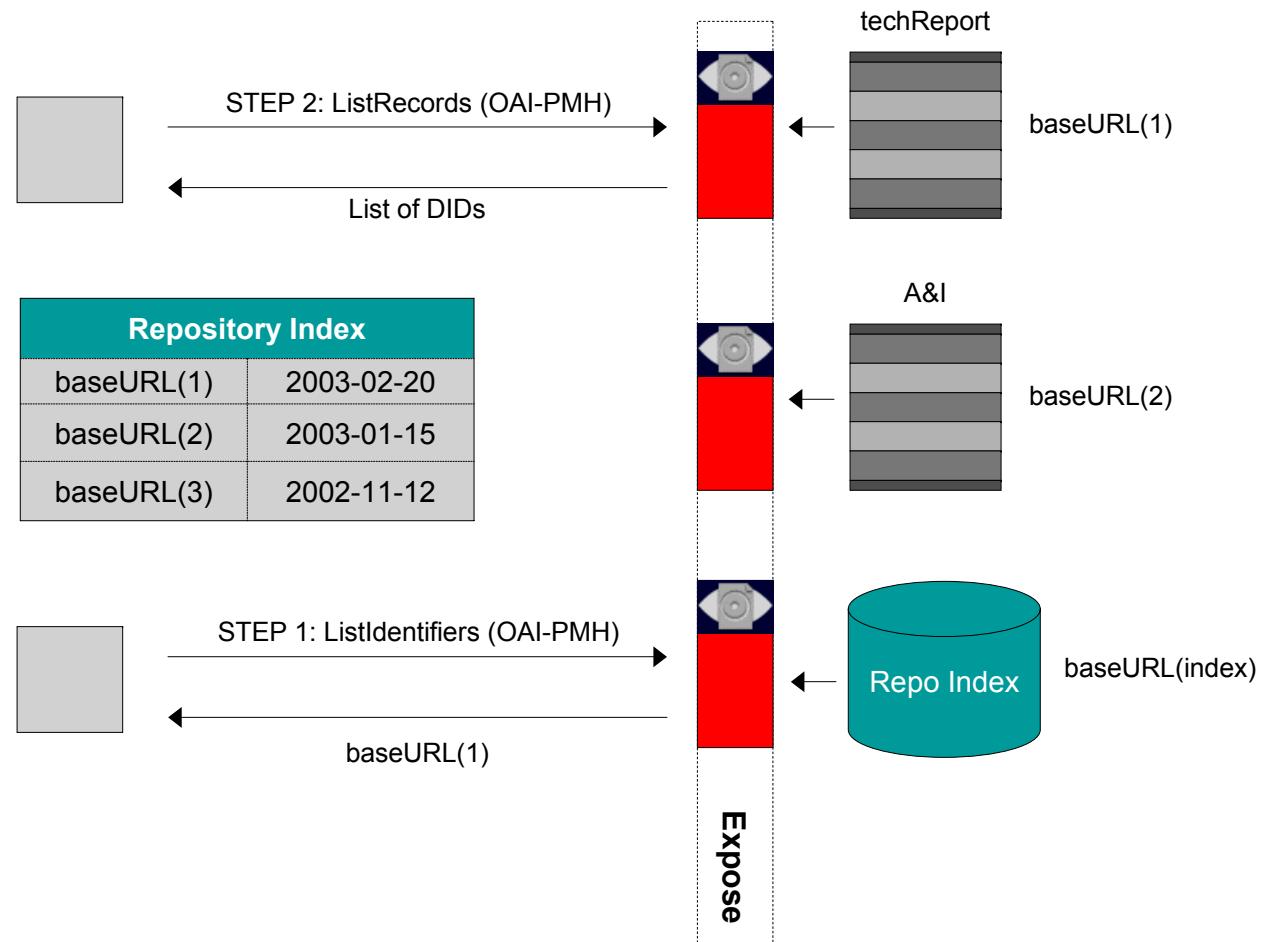


making DIDL documents accessible through the OAI-PMH

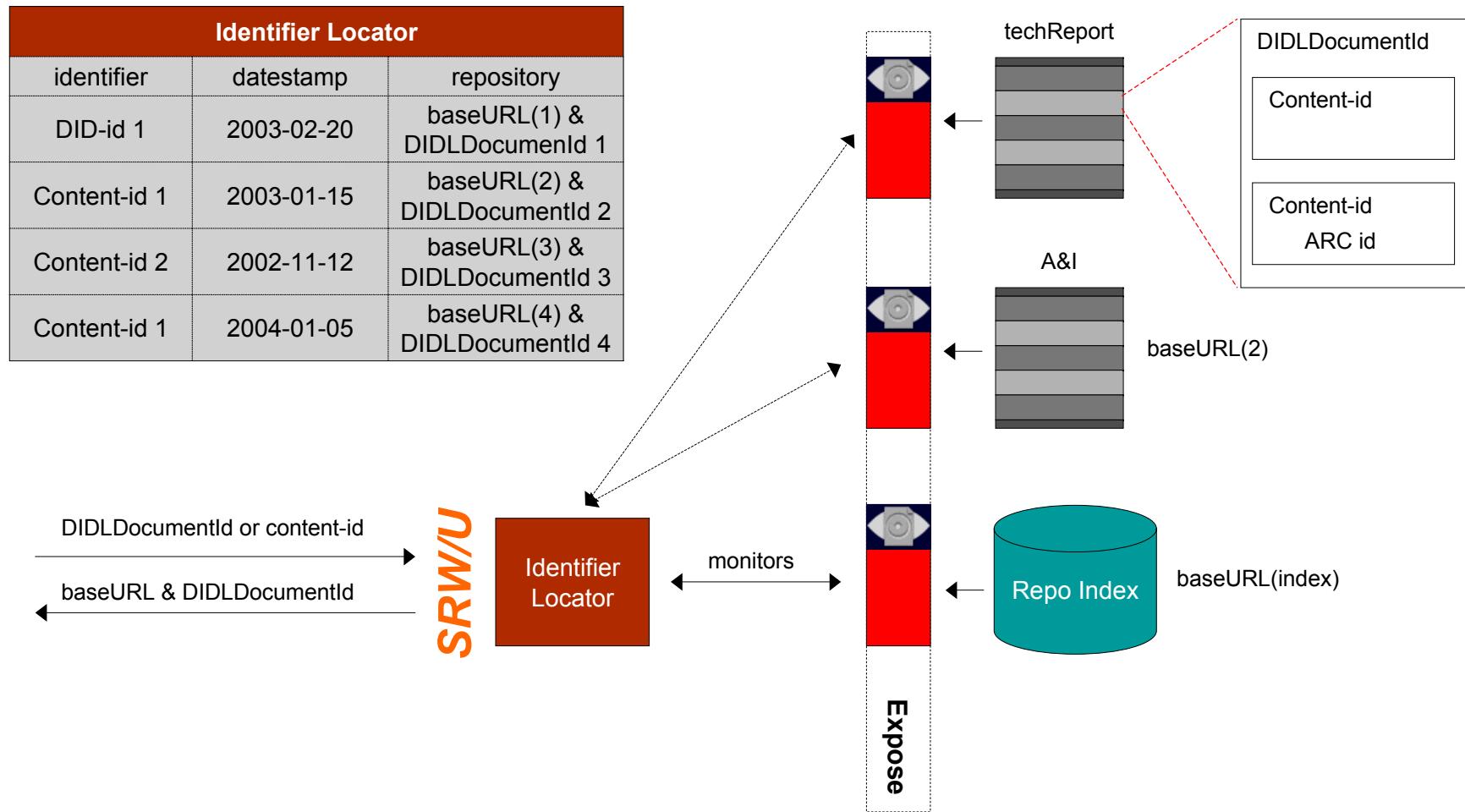




Repository Index: keeping track of OAI-PMH repositories

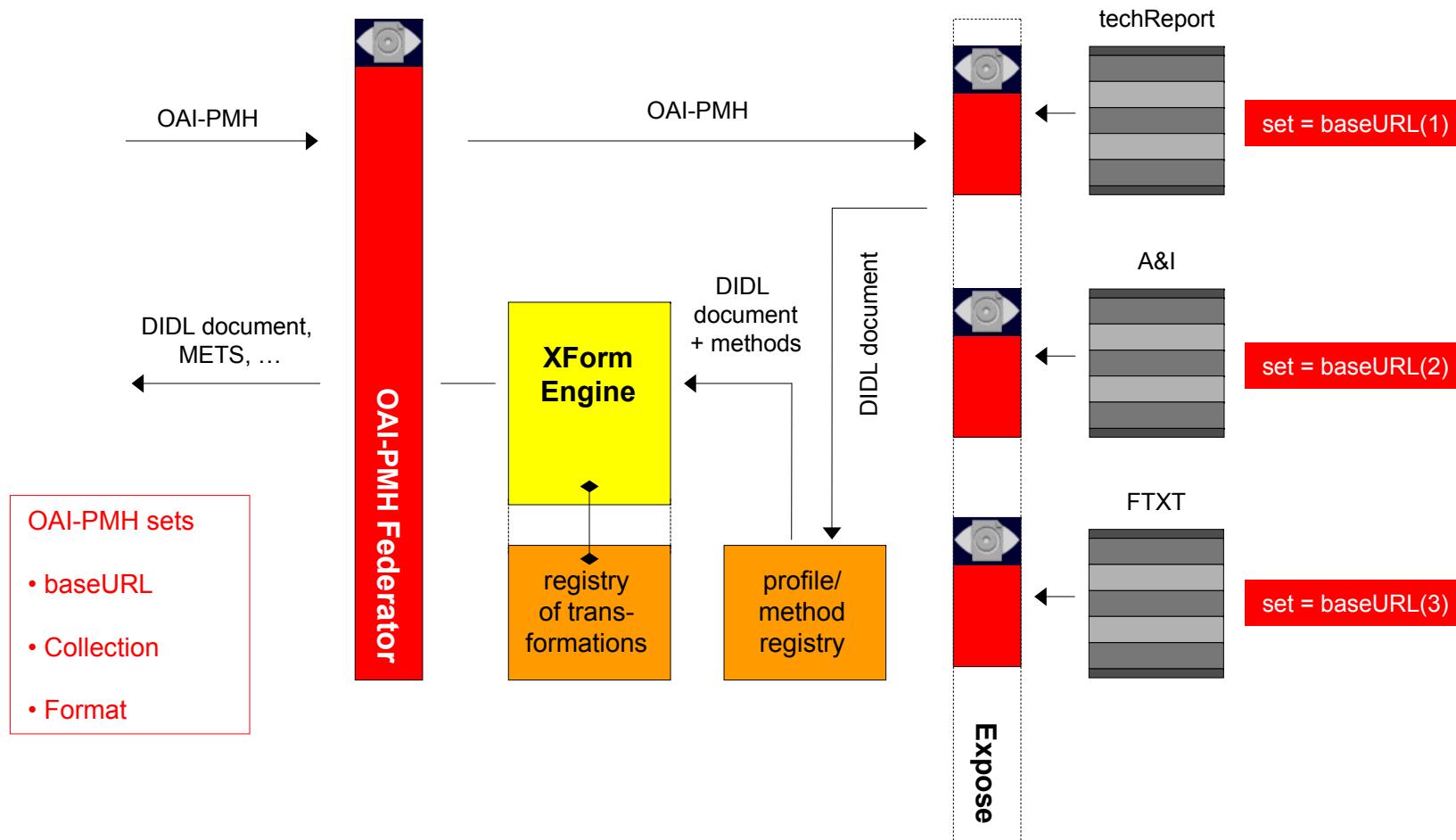


Identifier Locator: locating DIDL documents and Digital Objects



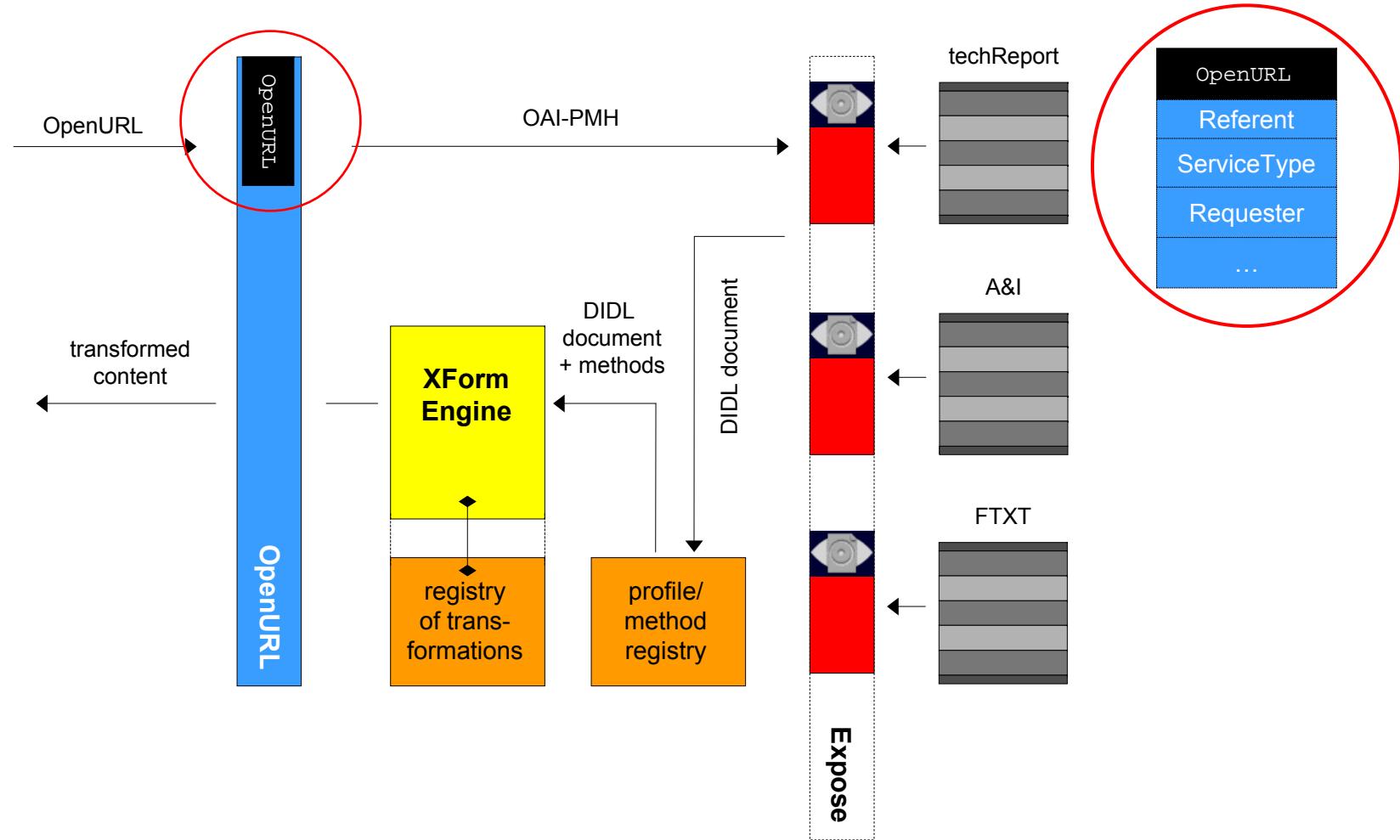


OAI-PMH Federator: single point of access to DIDL documents





OpenURL access to Items across repositories



OpenURL-based disseminations

- Disseminate DIDL documents, Digital Objects, constituent datastreams and transformations thereof
- Example:

```
http://gws.lanl.gov:9080/openurl-servlet/test?
& rfr_id=info:sid/library.lanl.gov
& url_ver=Z39.88-2004
& rft_id=info:lanl-repo/biosis/PREV196905076682
& svc_id=info:lanl-repo/svc/tomods.marc
```

OAI-PMH Federator & OpenURL Resolver

aDORe front-end	Interface standard	identifier	OAIS Access Type	# items in response
OAI-PMH Federator	OAI-PMH	Package Identifier	OAIS DIP	1 or more
OpenURL Resolver	NISO OpenURL	Content Identifier, Package Identifier (with XML ID fragment)	OAIS DIP & Result Set	1



readings

- Using MPEG-21 DIDL to Represent Complex Digital Objects in LANL
<http://www.dlib.org/dlib/november03/bekaert/11bekaert.html>
- Using MPEG-21 DIP and NISO OpenURL for the Dynamic Dissemination of Complex Digital Objects in LANL
<http://www.dlib.org/dlib/february04/bekaert/02bekaert.html>
- The multi-faceted use of the OAI-PMH in the LANL Repository
<http://lib-www.lanl.gov/~herbertv/papers/jcdl2004-submitted-draft.pdf>
- aDORe: a modular, standards-based Digital Object Repository.
<http://arxiv.org/abs/cs.DL/0502028>
- File-based storage of Digital Objects and constituent datastreams:
XMLtapes and Internet Archive ARC files.
<http://arxiv.org/abs/cs.DL/0503016>